Time to Discipline? Estimating the Risks and Impact of Public-School Discipline

Time to Discipline?

Estimating the Risks and Impact of Public-School Discipline

Stephen L. Hoffman

John B. Willett, chair
Daniel M. Koretz
Richard J. Murnane

A Thesis Presented to the Faculty
of the Graduate School of Education of Harvard University
in Partial Fulfillment of the Requirements
for the Degree of Doctor of Education

2016

Dedication

I dedicate this work to my wife, Tara Affolter, the first Ph.D. in our family. Her love and support has been a true source of strength. To my son, Jerry—still at high risk of disciplinary events—whose experience as a Black kid reminds us all how difficult it can truly be to just "make it" in the public schools. To my sister, Deborah Hoffman, the best principal I've ever known, who is still doing the gritty work of leading a school, now in her 20th year as a public-school principal. To my father, David Hoffman and my step-mother, Dr. Sylvia Granger, whose encouragement has been terrific, as I pursued this quixotic endeavor in my relative old age. To my friend and colleague George Theoharis, whose mentorship and collaboration has been crucial to my development as a scholar. To my advisor, John Willett, whose commitment to me never wavered. And finally, to Milt McPike, my first boss as a rookie assistant principal, now 18 years ago, whose passing is still mourned on the East side. A towering man with enormous faith in the people who served in the public schools, Milt presided as the principal of Madison East High School for 23 years with integrity, dignity and true character.

Acknowledgements

I am truly grateful for crucial design and methods assistance provided to me by my doctoral advisor, John Willett of the Harvard Graduate School of Education. He has worked tirelessly on drafts of all three of these papers, and his support has been immeasurable. But in addition, I value his friendship, kindness, and his steady belief in me. Professor Richard Murnane also provided substantial assistance to me, as I conceptualized and wrote these essays. In particular, his work with me on the first essay was pivotal. I also learned a great deal from him during my year-long research association with him—an apprenticeship that has shaped me greatly. Professor Daniel Koretz also provided vital critiques on my work throughout my doctoral career, and he has also been a valuable mentor to me—both as a researcher and as a teacher. Professor Katherine Masyn worked with me on early work for these essays, and her work on multiple-spells survival analysis has provided me with a great model to emulate. I also wish to acknowledge very helpful feedback on early conceptualizations of the first essay by Ben Castleman, North Cooc, Alejandro Ganimian, Alonso Sanchez, Rebecca Unterman, and two anonymous reviewers. For the second essay, I also appreciate the insightful comments of the assembled HGSE students and faculty made during a Quantitative Policy Analysis in Education luncheon/workshop. Similarly, I appreciate the helpful comments and suggestions provided to me by the attendees of the Educational Policy Colloquium at HGSE, when I presented an early draft of the third essay.

Table of Contents

Abstract

In the three essays in this thesis, I explore the effect of school discipline policies on the suspension of public-school students, in an urban setting.

In the first essay, using aggregate data, I investigate the effect of zero-tolerance disciplinary polices on secondary-school students. Capitalizing on a natural experiment, I used a "differences-in-differences" analytic approach to explore any benefit of a hypothesized deterrent effect and to estimate the impact of the abrupt expansion of zero-tolerance policies in one large urban school district. I found that Black students were *suspended* from school more often following the policy change, while suspensions of White students remained unchanged. In addition, *expulsions* from school, following the policy change, more than doubled for Black students, compared to only a small increase for White students.

In the second essay, and the same urban setting, I employed continuous-time survival analysis in a student-level event-history dataset to estimate the risk of middle-school students' first suspension of the school year. I found that this risk differed by three factors: (a) when the suspension occurred, (b) student grade-level, and (c) student race. At the beginning of the school year, this risk of first suspension for eighth-grade students was double the risk for sixth-grade students, although this difference diminished over time. Additionally, the risk for Black students was more than ten times the risk for White students.

In the third essay, I extended my work further, using repeated-spells survival analysis to describe the timing of suspensions over the duration of the students' entire middle-school careers. I found that—once a student had been suspended from middle

school for the first time—the median time until a second suspension was less than one school year, and the median time until a third suspension was about one semester. These risks also differed substantially by gender, race, and poverty level. The risk of a first suspension for boys was substantially higher than for girls. This risk was also higher for poor students than for non-poor students. However, the risks of *both* a first suspension and *subsequent* suspensions were substantially higher for Black students, compared to White students, even after controlling for differences in poverty among the groups.

Taken together, these analyses underscore disparities in school disciplinary practices, based on important student demographic characteristics, while providing an updated and more methodologically sound way of describing these effects.

Time to Discipline?

Estimating the Risks and Impact of Public-School Discipline

Introduction

Of the roughly 50 million children and young adults attending U.S. public schools, more than 6% are suspended out-of-school at least once during each school year. Employing the use of school exclusion—such as out-of-school suspension—has been a common practice for decades, although prior to the 1970s, such practices were almost entirely informal and discretionary. Additionally, disproportionate school disciplinary outcomes for students of color, particularly Black students, are pervasive in the United States, and the evidence of these disparities is overwhelming and well documented. While the persistent American "achievement gap" between Black and White students on myriad measures of academic achievement commands the focus of educators, policymakers, and researchers, enormous inequalities in school discipline between Black students and White students—a discipline gap— receive less policy attention.

For the ten years prior to beginning my doctoral career, I was part of the "system" of school discipline, working as a middle-school principal and a high-school assistant principal. On most days of the school year, I assigned formal discipline to students, following district procedures and the approach to discipline advocated by my superiors. And from my first day as an administrator, I was acutely aware of the racial dynamics in schools, particularly regarding the astounding disproportionality of school discipline. And so, in this thesis, I explore ways to unpack the phenomenon of school suspensions in three essays.

In the first essay, using aggregate data, I investigate the effect of zero-tolerance disciplinary polices on racial disparities in school discipline. I focused on secondary-school students attending one large, urban school district, and the school districts in the surrounding metropolitan area. Capitalizing on a natural experiment, I used a "differences-in-differences" analytic approach to explore any benefit of a hypothesized deterrent effect and to estimate the impact of the abrupt expansion of zero-tolerance policies in one large urban school district. I found that Black students were *suspended* from school more often following the policy change, while suspensions of White students remained unchanged. In addition, *expulsions* from school, following the policy change, more than doubled for Black students, compared to only a small increase for White students. This essay, subsequently published in *Educational Policy*, employed causal methodology that is rare in studies about school discipline.

Alas, four years ago, when I was still taking methodological coursework, I suffered an ischemic stroke. This set me back considerably, as my brain was thoroughly scrambled by this serious medical event. But if there were upsides to *surviving* this experience, perhaps one positive aspect was that it stimulated my interest in learning and using survival-analysis techniques as my main methodological framework for the second and third essays, as I recovered and continued my research.

Thus, in the second essay of my thesis, set in the same metropolitan area, I employed continuous-time survival analysis in a student-level event-history dataset to estimate the risk of middle-school students' first suspension of the school year. I found that this risk differed by three factors: (a) when the suspension occurred, (b) student grade-level, and (c) student race. In particular, I found that at the beginning of the school

year, this risk of first suspension for eighth-grade students was double the risk for sixth-grade students, although this difference, by grade level, diminished over time.

More importantly, I found that the risk of first suspension for Black students was more than *ten times* the risk of a first suspension for White students. The risk of first suspension for Hispanic students was lower than my estimates of risk for Black students but higher than my estimates for White students. And the risk of first suspension for Asian students was substantially lower than the risk for White students.

Beyond estimating the risk of a first suspension—risk that can also be estimated, to an extent, using cross-tabulations and regression techniques—I then expanded my use of survival-analysis techniques to describe the risk of repeated out-of-school suspensions. In the third essay, I focused again on middle-school students, using a confidential student-level dataset. Estimating the risk of suspension for a sample different from the second essay, I used repeated-spells discrete-time survival analysis to describe the timing of out-of-school suspensions for a cohort of students in one school district who entered 6^{th} grade together in the fall of 2007. I then followed this cohort of students over the entire sixth through eighth grade period, estimating differences in the level and shape of the risk function.

I found that—once a student had been suspended a first time—the median time until a second suspension was 170 school days. The median time until a third suspension was even shorter—only 98 school days. I also concluded that the risk of suspension differed substantially by gender, by race, and by students' level poverty, as measured by their eligibility for free or reduced-price lunch. Because the results of this analysis are complex, interacting spell, time and critical covariates, I also elected to display them

graphically, with a series of predicted hazard probability and survival probability functions.

I explored the effect of student gender first, and concluded that the risk of a first suspension agreed with other estimates of the difference in the suspension "rate". But I also found that, once a student was suspended, the risk of subsequent suspensions was equal, regardless of student gender. However, when I explored the effect of student race, my conclusions were different. The risk of first suspension for Black students was ten times higher than for White students. And furthermore, the risk of subsequent suspensions was also higher for Black students than for White students—controlling for a first suspension. I then explored the isolated effect of poverty, as measured by students' eligibility for free or reduced-price lunch. I concluded that the effects echo those that I found for gender: poor students had a relatively high risk of first suspension compared to non-poor students, but I found no difference in my estimates of risk between poor and non-poor students for subsequent suspensions.

Finally, in estimating the risk of suspension for all three critical covariates simultaneously, I found multifaceted results best explained graphically. The effect of poverty was highest for White female students, and lower for White males and for Black students. But the risk of suspension—both a first and subsequent suspensions—was substantially higher for Black students compared to White students, even after controlling for students' level of poverty.

Zero Benefit: Estimating the Effect of Zero-Tolerance Discipline Policies on Racial Disparities in School Discipline

Stephen L. Hoffman

Harvard Graduate School of Education

Zero Benefit: Estimating the Effect of Zero-Tolerance Discipline Policies on Racial Disparities in School Discipline

Disproportionate school disciplinary outcomes for students of color, particularly Black students, are pervasive in the United States, and the evidence of these disparities is overwhelming and well documented (Gregory, Skiba, & Noguera, 2010; McCarthy & Hoge, 1987; Nichols, 2004; Raffaele Mendez & Knoff, 2003; Skiba, Michael, Nardo, & Peterson, 2002; Townsend, 2000). While the persistent American "achievement gap" between Black and White students on myriad measures of academic achievement commands the focus of educators, policymakers, and researchers, enormous inequalities in school discipline between Black students and White students—a discipline gap—receive less policy attention. However, *Education Week* reported in 2010 that the federal government is investigating differences in disciplinary outcomes between White students and students of color, including the harshness of punishment and the disproportionate impact of "zero tolerance" disciplinary policies (Zehr, 2010).

Zero tolerance disciplinary policies warrant particular scrutiny, both because of the disparate impact on students of color, and because of questions regarding their effectiveness. An American Psychological Association (APA) Zero Tolerance Task Force recently concluded that the implementation of zero tolerance policies in the late 1990's and early 2000's did not improve school climate or school safety, and it may have exacerbated the discipline gap between White students and students of color (American Psychological Association, 2008). Asserting that "the time is right to end zero tolerance," LaMarche (2011) wrote in *Education Week* that zero tolerance policies have led to

suspension and expulsion rates at crisis proportions, denying students access to vital services, while failing to improve student behavior. The present study aims to examine specific evidence about the effects of the expansion of zero tolerance discipline policy on school suspension and expulsion rates for both Black students and White students, as well as on the rate of alcohol, drug, and weapons violations in an urban school district.

Background and Context

Racial disparities in school discipline in U.S. schools have been documented in scholarly articles for decades. McCarthy and Hoge (1987) reviewed literature from the 1960's through the 1980's that documented Black students being suspended from school or otherwise disciplined at rates more than three times that of White students. In their investigation of a mid-Atlantic city in the 1970's, they concluded that significant disparities in punishment were not reasonably explained by differences in student misbehavior, and they noted a high degree of subjectivity among school authorities in decisions about school discipline. Bowditch (1993) documented how school suspensions for Black students were seemingly disproportionate to the nature of the violations, and that school staff frequently used student transfers or the involuntary dropping of Black students as disciplinary tools. Raffaele Mendez and Knoff (2003) noted that out-of-school suspensions in the U.S. increased during the 1990's, and that 7th, 8th, and 9th grade students, particularly minority students, were most frequently suspended. They reported that Black students in the 1990's were suspended, on average, approximately 2.3 times more often than White students, although they noted some school districts where the suspension rate for Black students was as high as 22 times the rate for White students. Gregory and Mosely (2004) studied racial disparities in achievement and discipline at a

large, diverse, urban high school, documenting a discipline gap for both within-school sanctions and suspension, documenting that both Black students represented approximately 37% of the student population but accounted for 80% of students sent to On-Campus Suspension and 68% of out of school suspensions.

Recent data from the National Center for Educational Statistics (NCES) indicates that racial disparities in school discipline have persisted and are arguably worsening during the last decade. School suspension and expulsion continue to be common forms of punishment in American schools. More than 3.3 million American students were suspended and over 102,000 were expelled from school in 2006[1] (NCES, 2009). Furthermore, the racial/ethnic distribution of these suspensions and expulsions reveal stark disparities: 15% of Black students, 6.8% of Hispanic students, 4.8% of White students, and 2.7% of Asian students were suspended from school. Using the Parent and Family Involvement in Education Survey in 1999, 2003, and 2007, the NCES (2012) estimates that the percentage of Black public school high school students who had ever been suspended rose from 37% in 1999 to 49% in 2007, compared to the White rate of 18.2% in 1999 and 17.7% in 2007. Similarly, the estimated percentage of Black students who had ever been expelled from school rose from 6.5% in 1999 to 10.3% in 2007, while the rate for White students dropped from 1.8% in 1999 to 1.1% in 2007.

[1] The National Center for Educational Statistics and the Office of Civil Rights define suspension as an out-of-school suspension, during which a student is excluded from school for disciplinary reasons for 1 school day or longer; it does not include students who served their suspension in the school. Expulsion is defined as the exclusion of a student from school for disciplinary reasons that results in the student's removal from school attendance rolls or that meets the criteria for expulsion as defined by the appropriate state or local school authority. For both suspensions and expulsions, students are counted only once, but may appear in both categories.

This widening of the discipline gap occurred during a period of significant expansion of zero tolerance discipline policies. Zero tolerance policies are "defined as a school or district policy that mandates predetermined consequence/s or punishments for specific offenses" (U.S. Department of Education, 1998, p. 18). Federal influence on school discipline policy and zero tolerance policies in particular originated with the Gun-Free Schools Act of 1994, which directed states to pass legislation mandating the automatic expulsion of students from public schools for possessing a weapon (Sughrue, 2003). However in many schools, the concept of zero tolerance has since evolved to include the automatic suspension or expulsion of students for an expanded list of offenses, including alcohol and drug violations, physical assault and fighting, criminal damage to property, and committing multiple violations in the same school year (a closely related "three strikes" disciplinary policy).

Researchers and advocates who express concern about zero tolerance disciplinary policies acknowledge that school safety and the protection of students and staff from violence and illegal drugs is vital, but question the effectiveness and fairness of such policies (Sughrue, 2003; Skiba and Rausch, 2006). While a get-tough attitude about school discipline may seem like a sensible approach, Skiba and Rausch (2006) report that zero tolerance discipline policies are associated with poorer school climate, lower student achievement, higher rates drop-out rates, and that increased reliance on suspension and expulsion for maintaining school climate and safety is likely to exacerbate racial disparities already present between Black and White students.

In 2005, the American Psychological Association (APA) commissioned a task force to explore the impact of zero tolerance discipline policies in elementary and

secondary schools. While acknowledging that safe and disciplined schools are a vital policy goal, the Task Force found little evidence to support the basic assumptions of a zero tolerance approach: that the certainty and seriousness of punishment will have a deterrent effect on students; that removing severely disruptive students will deter other students from behaving in a similar manner; and that removing offenders will improve school climate. Instead, the Task Force concluded that the available evidence tended to indicate that suspending students predicts more future misbehavior and that schools with higher rates of suspension and expulsion have poorer climate (American Psychological Association, 2008).

As to the impact of zero tolerance policies specifically on the discipline gap, the APA Task Force notes that by decreasing the subjectivity of decision-making regarding discipline, perhaps such policies would reduce some bias and be fairer to students who traditionally have been subjected to harsher discipline (American Psychological Association, 2008). However, critical race theorists in education (e.g. Ladson-Billings, 1998; DeCuir & Dixson, 2004; Gillborn, 2005) argue against the notion that policies can be racially neutral in our present school system, noting that policies and practices subtly privilege White students while casting Black students as deficient and in need of "fixing." Gillborn (2005) notes how policy-makers mistakenly envision education policy as consistently making at least incremental progress, and they frequently assume that a new policy (like zero tolerance) is naturally an improvement that can escape the racism of previous policies. Yet, as DeCuir and Dixson (2004) note, decreased subjectivity and notions of colorblind policies fail to consider persistent racism and how policies that subtly privilege White students might interact with a policy like zero tolerance. As

Casella (2003) argues, "punishment negatively affects those who are already negatively affected by poverty, racism, academic failure, and other realities" (p. 879). What appears to be "neutral" policy that reduces some subjectivity of interpretation by school authorities might still be associated with increased racial disparities in discipline outcomes.

In summary, despite the attention and alarm raised during the several decades about the discipline gap, racial disparities in school suspension and expulsion worsened considerably between 1999 and 2007 (NCES, 2012), and this tragic phenomenon roughly coincided with the expansion of zero tolerance discipline policies in various states and districts. The APA Task Force specifically called for researchers to "conduct systematic efficacy research including quasi-experimental and randomized designs to compare outcomes of programs with and without zero tolerance policies and practices" (American Psychological Association, 2008, p. 859). The present study is a response to that call. Using a quasi-experimental design, this study exploits a school district policy discontinuity—the abrupt expansion of zero tolerance discipline policy in a mid-sized urban school district (here-to-for referred to as the "District")—to estimate the causal impact of zero tolerance discipline policies on racial disparities in disciplinary outcomes.

Research Questions and Hypotheses

In this study, I estimate the effect of the expansion of this zero tolerance discipline policy on two different discipline outcomes: racial differences in the percentage of students recommended for expulsion from the District; and racial differences in the proportion of days that all students in District secondary schools were suspended from school for any reason, including those students who were disciplined for

less serious infractions. Beyond the effect of the change in disciplinary policy on the relatively small proportion of students who commit serious offenses and are recommended for expulsion, I also sought evidence about the effects of expanded zero tolerance on students who are *not* recommended for expulsion. If zero tolerance policies have a deterrent effect for the larger population of students, and if an improved climate for learning is expected, then implementation should cause a decline in the incidents leading to school discipline—particularly school suspension. However, zero tolerance may also signal to school staff that they need to be more strict assigning discipline to students *for all offenses*, not just those that are the most serious. And if zero tolerance discipline policies influence staff to administer harsher punishments in general, will this affect Black students and White students differently? Consequently, I pose the following research questions:

1. **Did expanding the zero tolerance disciplinary policy significantly widen racial disparities in the percentage of students recommended for expulsion?**
2. **Did expanding the zero tolerance disciplinary policy affect the proportion of days that Black students and White students were suspended from school?**

I hypothesized that the expansion of zero tolerance disciplinary policy exacerbated the already substantial disparities in expulsions and suspensions between Black students and White students. Defining a larger number of behavioral offenses as cause for recommending the expulsion of students would increase the number of students recommended, and that increase would disproportionately affect Black students. Additionally, while some students may have altered their behavior in response to the

advertised changes in discipline policy, I hypothesized that overall suspension rates would increase, again disproportionately affecting Black students.

Methods

The site of the present study is a mid-sized urban school district serving more than 20,000 students. The District touts the high quality of its schools on the district website, noting that it well exceeds national and state averages for the rate that students pass Advanced Placement exams and are named National Merit Scholars. It is also a diverse school district: 50% White, 24% Black, 15% Hispanic, and 10% Asian, and is committed through its mission statement to "embracing the full richness and diversity of our community." Unfortunately, as is all too common, the experience of students in schools differs significantly by race, particularly in who experiences school discipline and school removal. During the 2009-10 school year, more than 33% of the District's approximately 3,000 Black secondary school students were suspended from school at least once, compared to 5% of the District's 6,500 White secondary school students—a racial disparity in the percentage of students suspended of more than 6 to 1.

A Natural Experiment

In September 2007, the District instituted a significant and unadvertised policy change regarding student discipline, substantially expanding the list of offenses subject to a zero tolerance mandate. This sharp discontinuity in school discipline policy provides an opportunity to study the effects of zero tolerance discipline on racial disparities in school discipline outcomes. Initiated by the school board and school district administration, and introduced into the student code of conduct at the beginning of the 2007-08 school year, this policy change mandated the use of an "aggravating factors analysis" by secondary

school principals for serious violations of school rules. Many offenses that had previously been dealt with at the school level now required that principals suspend the student for five days and recommend the expulsion of that student to the superintendent of schools.

School districts typically expel students for very serious offenses, such as possession/use of a weapon, physical assault of staff, or selling illegal drugs at school. Like many other school districts, the District had codified that students who commit these serious offenses must be recommended for expulsion from school. Other serious offenses, like fighting, physical assault, property damage, and bomb threats had also been cause for suspension; and until September 2007 these offenses *may* have led to a recommendation for expulsion. Similarly, repeated serious violations of school rules also *may* have led to a recommendation for expulsion. However, beginning in September 2007, the Student Code of Conduct was changed, and principals were now *required* to recommend the expulsion of secondary students who commit a serious violation of school rules *if* one of several "aggravating factors" was determined to be present, including serious bodily injury, significant property damage, arrest for a Class A Misdemeanor or higher, and/or a significant loss of instructional time. Furthermore, under the expanded zero tolerance policy, principals were also required to recommend the expulsion of secondary students who committed three separate, serious violations within the same school year (fighting, stealing, and using alcohol, for example). This discipline policy applied to all District secondary students, which were defined as students in grades 6 through 12. In this analysis, I focus on expulsion *recommendations*, as opposed to actual expulsions. The process of actually expelling a student from the District involves many layers of administrative process, an analysis of which would

require access to private student data and the confidential records of expulsion hearings and closed Board of Education meetings. Such an analysis is beyond the scope of this essay.[2]

My analysis of the minutes of Board of Education meetings in 2007 indicated that district-level administrators authored the "aggravating factors analysis" during the summer of 2007, in an attempt to bring consistency to the recommendation for expulsion process across the district. At the meeting in August 2007 where it was approved by the Board of Education, the policy was reviewed in detail. Nothing in the minutes indicates that the policy was a response to increased discipline infractions in school, or to any public discussion of school discipline. Nor is there evidence that other school districts in the area made any similar changes in discipline policy. The published minutes from a Board of Education meeting confirm that the policy change was an attempt to remove subjectivity, and possibly to reduce the number of expulsions. "Adding the aggravating factors removes all the discretion and objectifies the process. It will be far more consistent and may reduce the number of expulsions" (School Board Minutes). This

[2] An expulsion recommendation is a very serious action taken by the school district, which begins in almost all cases with a 10-day suspension from school, a series of legal notices from school district lawyer, and an educational records review. This review process is designed to ascertain whether a student may have an undiagnosed educational disability "If at the end of the process the student is suspected to have a disability, the district conducts a special education evaluation of the student. The expulsion process is postponed during that evaluation and the student receives Off-Campus instruction, which is provided for 2 hours per day at a neutral site (library, community center, etc). If the student is found to have a disability and the Manifestation Determination concludes that any discovered disability "was a substantial cause of the incident, then the case is dismissed. If not, then it proceeds through the expulsion process." If an expulsion recommendation continues beyond a Manifestation Determination, there is a "trial" presided over by a neutral hearing officer contracted by the school district, with the district represented by legal counsel. The decision of this hearing officer is forwarded to the Board of Education, for their review and ultimate disposition, in closed session.

abrupt expansion of zero tolerance disciplinary policy by the school board, beginning in September 2007 was unanticipated by school staff, students, and parents. Therefore, I assert that this policy change constituted an exogenous "treatment" of secondary school students in the District (Murnane & Willett, 2011).

Recommendations for Expulsion

I use two separate sources of data to address my research questions. The first dataset is a compilation of the number of students recommended for expulsion from the District from 2005-06 through 2008-09, disaggregated by race, as reported in the District's "Disciplinary Options Expulsion Data Summary." Students in this dataset were assigned to one of five race/ethnicity categories with which the student most identified: Asian (including Pacific Islander), Black (not Hispanic), Hispanic (all races), Native American (American Indian or Alaska Native) or White (not Hispanic). Combining this information with enrollment information available from the State Department of Education allowed me to calculate the percentage of secondary students of each race/ethnicity recommended for expulsion in the two school years before the policy discontinuity and the two school years after the policy was implemented.

Proportion of Days Suspended

The second dataset is compiled from data available from the State Department of Education. School suspension rates, including the rates for the District reported at the beginning of this paper, are often reported as percentages of students suspended in a given year. My analysis uses a less common measure, the proportion of days lost due to suspension, in order to capture both the overall rate of behavior leading to school

suspension, as well as the response of school staff to the severity of the offenses committed.

The outcome variable of interest in this dataset is the proportion of possible school days students in each school were suspended, in the school years before and after the policy change. Proportion of days suspended is calculated by totaling the number of school days that students in a given school, of a given race, were suspended, and then dividing that number by the possible number of school days that students of a given race could attend school (e.g. 200 White students enrolled for the entire school year of 180 days = 36,000, which is the number of possible school days for White students). This variable is recorded twice for each school in each year: once as the proportion of days that Black students are suspended from a school in a particular year, and also as the proportion of days that White students are suspended from the same school in the same year.[3] Values of this proportion range from zero, in three middle schools that suspended no White students during one entire school year, up to 0.019 in a middle school where Black students were suspended for nearly 2% of possible school days—an average of more than 3 days of suspension for every Black student in the school that year.

Additionally, the percentage of economically disadvantaged students was included for use as a covariate. Time is included as a continuous variable measured in school years and centered at the policy continuity. A dichotomous variable indicating whether the expanded zero tolerance policy was in effect served as the predictor of interest.

---

[3] Suspension data regarding students of other races/ethnicities is only sporadically available publicly, because data for racial groups with smaller numbers of students are frequently suppressed for student privacy.

The sample is comprised of 37 secondary schools: 15 District schools and 22 comparison schools. All of the traditional secondary schools in the District are included in the sample: 4 comprehensive high schools and 11 middle schools.[4] In order to provide evidence about secular trends in suspension rates during this time period for similar schools that were not subject to the policy change, 22 comparison schools from the area are included in the dataset. This sample of comparison schools from the same county as the District, as well as the secondary schools in a school district in the same athletic conference is comprised of 9 medium and large high schools, and the 13 middle schools that feed into them.[5] All of these schools operate under the same state laws as the District, but none of the school districts governing these comparison schools utilized a zero tolerance approach to discipline. I found no evidence of substantial discipline policy changes in the comparison schools during the time frame of this study.[6] Consequently, these schools serve as reasonable comparison schools to use in estimating trends in school suspensions for District schools, had the zero tolerance policy not been expanded.

The outcome of interest, the percentage of days students in a school were suspended, is positively skewed for White students but not for Black students.

---

[4] I excluded four small District alternative programs from my analysis of suspension data.
[5] One of the neighboring high schools in the county which was large enough to consider including in the sample was not included, because data on the number of suspension days for Black students was suppressed for two consecutive years, due to the very small number of Black students in the school. One other school district in the same conference athletic conference also implemented a zero tolerance discipline policy, and was not included in the estimates of the secular trend.
[6] I reviewed the publicly available student codes of conduct for the comparison school districts. None of the available records showed any evidence of substantial discipline policy changes, nor did most of the school districts expel more than a handful of students per year—indeed if any pattern was evident, it was that expulsion rates were declining in the comparison school districts.

Furthermore, these proportions are perhaps best conceptualized as a count of events (suspension days) accumulated during a school year by students attending school. Analysis of proportions and counts using common regression techniques can be problematic. Count variables are necessarily discrete and positive, and the distributions are frequently skewed. Thus, the assumption of normal distribution of errors is not tenable, and the conditional mean structure should be constrained to be positive (Allison, 2009; DeMaris, 2004).

One technique more appropriate for modeling count data is a negative binomial regression model, which assumes that the event of interest (days of suspension from school) is a count resulting from an underlying continuous process, and that the rate of occurrence is governed by a negative binomial distribution (DeMaris, 2004). The negative binomial distribution is a probability distribution frequently used for modeling the probability of success or failure over a series of independent, and identically distributed trials. In this particular case, each day of school that students are enrolled is modeled as a "trial", with days suspended modeled as "failures."[7]

For analyzing counts structured in panel data such as this dataset, Allison (2009) recommends estimating a negative binomial regression model using the **nbreg** command in Stata (StataCorp, 2011), including dummy variables for each school and correcting standard errors using the outer product of the gradient option. In this model, the number of days students are suspended is a function of the dichotomous main effect of whether the expanded zero tolerance policy is in force *(ZeroTolerance),* the dichotomous race

[7] Several other models yielded similar results, including fixed-effects regression using a logit-transform of the outcome variable, as well as fixed-effects regression of the raw percentages.

indicator variable *(Black)*, the interaction of *Black* and *ZeroTolerance,* school year *(Year)*, and the interaction of *Black* and *Year*.

I also included the time-varying percentage of economically disadvantaged students in the school*,* the number of possible attendance days for students of a particular race, and school dummy variables as fixed effects. This regression formulation of the difference-in-differences strategy allows me to include the time-varying school-level covariates (Angrist & Pischke, 2009), and to include the data from both District schools and comparison schools in the calculation of the underlying secular trend of school suspensions. The first parameter of interest is the causal effect of expanding the zero tolerance discipline policy on the number of suspension days for White students. The second parameter of interest is the additional causal effect of expanding the zero tolerance discipline policy on the number of suspension days for Black students, beyond the effect for White students.

Results

The Proportion of Students Recommended for Expulsion.

Did the expansion of the zero tolerance discipline policy increase the percentage of students recommended for expulsion? An examination of the descriptive statistics for expulsion recommendations during this time period reveals that the expanded zero tolerance policy did have that effect, and the effect on the population of Black students in the District is much greater than the effect on students of any other race or ethnicity. Nearly 200 secondary students were recommended for expulsion in the two years immediately before the expanded zero tolerance policy was implemented. This total nearly doubled to 380 students in the two years following the expansion of zero

tolerance, an increase in the percentage of students recommended for expulsion from 0.72% to 1.46% of secondary school students. To illustrate these findings, I present a plot showing the percentage of Black, Hispanic, and White secondary students recommended for expulsion from the District during the two school years before and the two school years after the expansion of the zero tolerance discipline policy in Figure 1. The percentage of White secondary students recommended for expulsion increased from 0.3% before the policy change to 0.5% after the policy change. For Hispanic students, the percent of students recommended for expulsion increased from 0.8% to 1.0% of secondary students. The already high percentage of Black secondary students recommended for expulsion from the District more than doubled from 2.2% to 4.5%.

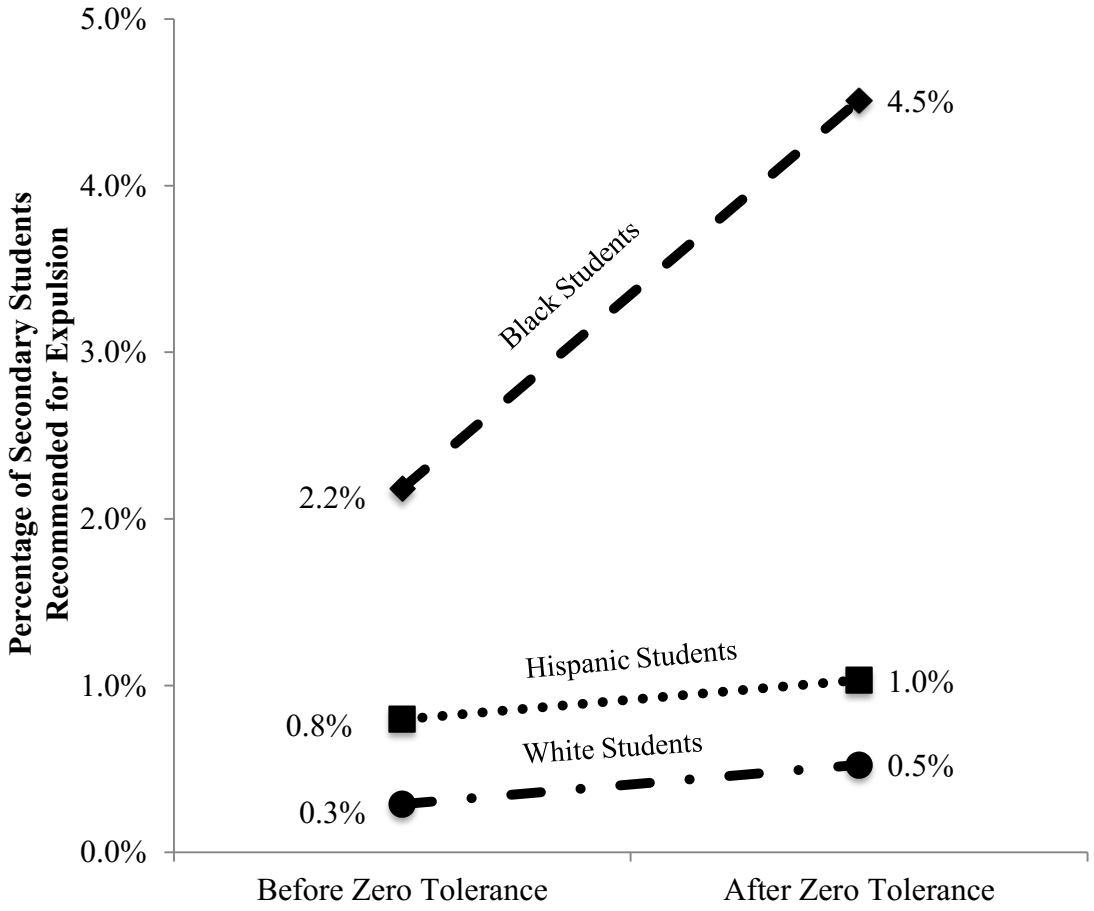

Figure 1. Percentage of Black, Hispanic, and White secondary students recommended for expulsion from the District during the two school years before and the two school years after the expansion of the zero tolerance discipline policy.

The Proportion of Days Students in a School are Suspended

Does expanding the zero tolerance disciplinary policy affect the percentage of days that students are suspended from school? The results of this analysis are mixed. The fitted percentage of days that Black students were suspended from school *increased* under the expanded zero tolerance policy, but the proportion of days that White students were suspended from school decreased a statistically non-significant amount.

To illustrate these effects, I present Figure 2, a plot of the fitted values of the (de-transformed) proportion of days suspended versus year (converted into percentages), for Black students and White students, at a prototypical District secondary school. Notice that the largest difference is between the proportion of days suspended for Black students and White students, regardless of the impact of the zero tolerance discipline policy. On average, Black students are suspended for about 0.7% of possible school days (1.25 school days per student, per year), while White students are suspended for approximately 0.1% of possible school days (0.18 school days per student, per year).

In Figure 2, I also illustrate the causal effects of the expansion of the zero tolerance discipline policy. For White students the proportion of days suspended remained virtually unchanged (a statistically non-significant effect of *zerotol*) at approximately 0.1% of possible school days. However, for Black students, the expansion of the zero tolerance discipline policy increased the fitted percentage of days suspended from 0.66% of possible school days to 0.76% of possible school days—an increase in the predicted number of days Black students in the District were suspended from school of nearly 0.2 days per Black student, in the 2007-08 school year, the first school year after the expansion of zero tolerance.

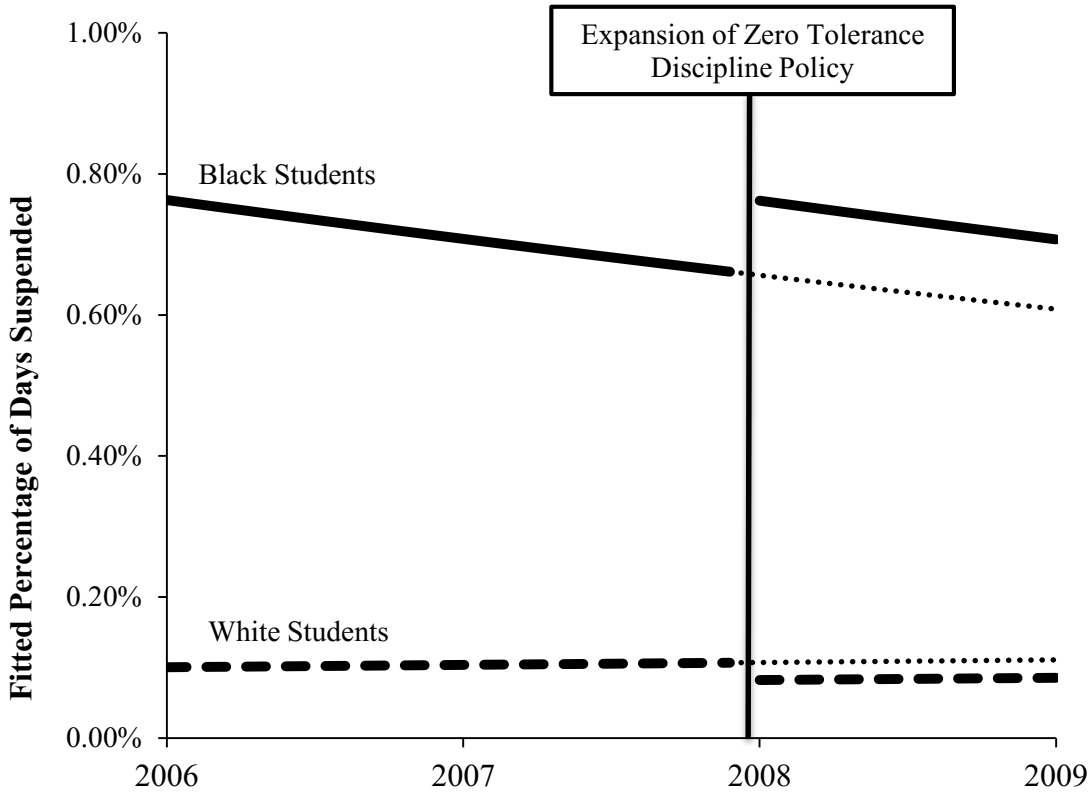

Figure 2. Plot of the fitted percentage of days suspended versus school year (where the year listed is the *end year* of the school year), for Black students and White students in a prototypical secondary school in the District; and illustrating the effect of the expansion of zero tolerance discipline policy, and the estimated secular trend for schools not subject to the expansion of zero tolerance discipline policy (school fixed-effects, $n_{schools}$=37, n_{years}=4).

Discussion

The expansion of the zero tolerance discipline policy in the District at the beginning of the 2007-08 school year led to a substantial increase in the percentage of Black secondary students being recommended for expulsion. This increase, from 2.2% of students before the policy change to 4.5% following the expansion of zero tolerance, resulted in the recommendation for expulsion of approximately 70 more Black students per year than would have been expected had the policy not been implemented. Additionally, approximately 20 more students of other races/ethnicities were recommended for expulsion per year than would have been expected. Although less than a quarter of the secondary students in the District are Black, they comprised about three-quarters of the increase in recommendations for expulsion under the expanded zero tolerance discipline policy. Clearly, racial disparities in rates of recommendation for expulsion were exacerbated under the expanded zero tolerance policy.

Did expanding zero tolerance affect the rest of the population of secondary students, perhaps by providing a deterrent effect, or by creating safer environments, more conducive to learning? Here again, my analysis demonstrates that expanding zero tolerance exacerbated already severe racial disparities in school disciplinary outcomes. I estimate that the percentage of days that White students were suspended from school was virtually unaffected by the expansion of zero tolerance, holding nearly steady at approximately 0.1% of possible days of attendance. However, for Black students, I estimate that the expansion of zero tolerance caused an increase in the percentage of days suspended from of approximately 0.1 percentage points—from 0.66% to 0.76%—for the 2007-08 school year. This increase in the percentage of days suspended is approximately

an additional 570 days of instruction lost to suspension for Black secondary students in the District.

Did expanding the zero tolerance policy make District schools safer? Again, my analysis of available data indicates that it did not. The State Department of Education collected information on the types of offenses students are suspended for, dividing them into weapon and drug violations, versus other violations. During the time period under study, District secondary students were suspended for approximately 20 weapon or drug violations per 1,000 students each year. To analyze the effect of expanding zero tolerance policies on type of offense, I found suggestive evidence of a significant *increase* in weapon and drug related offenses of 4.0 per thousand (t=1.72, p=.09) following the policy change. Thus, I did not detect an increase in school safety as a result of the expansion of zero tolerance discipline policy. Rather, I detected the opposite effect: a marginally significant increase in weapon and drug-related offenses in District schools, while Black secondary students, already suspended for more than 4,200 days in 2006-07 school year, lost approximately 570 more days of instruction due to the policy change.

Conclusion

My analysis of the expansion of a zero tolerance policy in a diverse urban school district supports LaMarche's (2011) assertion that the time is *indeed* right to end zero tolerance policies in America's public schools. The practice of mandating pre-determined disciplinary consequences for students does not serve as a deterrent for students, and it does not make schools safer. Furthermore, zero tolerance policies have an especially harsh impact on Black students, exacerbating already severe disparities in school discipline between Black students and White students. In the District, this discipline

policy applies equally to all secondary students—including students as young as 11 years old. Denying students, particularly young students, access to schools, and the related counseling and social work services that schools provide will not cause students to change their behavior. And abdicating responsibility for providing a free and appropriate public education for students who behave badly does not serve the public interest. Rather, it kicks the can down the road to future public agencies that will end up dealing with citizens whose education has been significantly disrupted in the name of "consistency" and "get tough" and "no excuses."

Nearly three years after it expanded the zero tolerance discipline policy, the District Board of Education authorized the implementation of a program to serve some expelled students. The administration noted in a report to the Board that "[c]oncerns have been raised by members of the Board of Education…staff and the community about the zero tolerance model, lack of services to expelled students and the significant disruption caused in the lives of these students, families and neighborhoods when students are expelled." Presented as an exhibit in Board minutes for the meeting proposing the new program was an article in *District Administration* outlining replacing zero tolerance policies with a restorative justice approach. Schacter (2010) describes how the school district in Denver CO, discarded its zero tolerance discipline policy in favor of positive behavior support and restorative justice practices, after receiving input from community stakeholders, the police, and the district attorney's office. Ending zero tolerance, in favor of proactive and compassionate approaches to discipline policy, is an important part of solving the "discipline gap" in American schools, and providing all students in the community with the skills and habits necessary for a successful life.

References

Alison, P. D. (2009). *Fixed effects regression model.* Thousand Oaks, CA: SAGE Publications, Inc.

American Psychological Association Zero Tolerance Task Force (2008). Are zero tolerance policies effective in the schools? An evidentiary review and recommendations. *American Psychologist, 63*(9), 852 – 862.

Angrist, J. D., and Pischke, J. (2009). *Mostly harmless econometrics: An empiricist's companion.* Princeton, NJ: Princeton University Press.

DeCuir, J. T., and Dixson, A. D. (2004). "So when it comes out, they aren't that surprised that it is there": Using critical race theory as a tool of analysis of race and racism in education. *Educational Researcher, 33*(5), 26-31.

DeMaris, A. (2004). *Regression with social data: Modeling continuous and limited response variables.* Hoboken, NJ: John Wiley & Sons, Inc.

Gillborn, D. (2005). Education policy as an act of white supremacy: Whiteness, critical race theory and education reform. *Journal of Education Policy, 20*(4), 485-505.

Gregory, A., and Mosely, P. M. (2004). The discipline gap: Teachers' views on the over-representation of African American students in the discipline system. *Equity & Excellence in Education, 37*, 18-30.

Gregory, A., Skiba, R., & Noguera, P. (2010). The achievement gap and the discipline gap: Two sides of the same coin? *Educational Researcher, 39*(1), 59-65.

Ladson-Billings, G. (1998). Just what is critical race theory and what's it doing in a nice field like education? *International Journal of Qualitative Studies in Education, 11*(1), 7-24.

LaMarche, G. (2011) The time is right to end 'zero tolerance.' *Education Week.* Published in Print: April 6, 2011. Retrieved from: http://www.edweek.org/ew/articles/2011/04/06/27lamarche.h30.html?r=901966954

McCarthy, J. D., and Hoge, D. R. (1987). The social construction of school punishment: Racial disadvantage out of universalistic process. *Social Forces, 65*(4), 1101-1120.

Murnane, R. J., & Willett, J. B. (2011). *Methods matter: Improving causal inference in educational and social science research.* New York, NY: Oxford University Press.

National Center for Educational Statistics (2009). School Characteristics and Climate. *Contexts of Elementary and Secondary Education.* Retrieved from http://nces.ed.gov/programs/coe/2009/section4/table-sdi-1.asp

National Center for Educational Statistics (2012). *America's youth: Transitions to adulthood.* Retrieved from http://nces.ed.gov/pubs2012/2012026.

Nichols, J. D. (2004). An exploration of discipline and suspension data. *The Journal of Negro Education, 73*(4), 408-423.

Raffaele Mendez, L. M., & Knoff, H. M. (2003). Who gets suspended from school and why: A demographic analysis of schools and disciplinary infractions in a large school district. *Education and Treatment of Children, 26*(1), 30-51.

Schacter, R. (2010). Discipline gets the boot. *District Administration*, January 2010. Retrieved from: http://www.districtadministration.com/viewarticle.aspx?articleid=2262.

Skiba, R. J., Michael, R. S., Nardo, A. C., & Peterson, R. L. (2002). The color of discipline: Sources of racial and gender disproportionality in school punishment. *The Urban Review, 34*(4), 317-342.

Skiba, R. J., & Rausch, M. K. (2006). Zero tolerance, suspension, and expulsion: Questions of equity and effectiveness. In Evertson, C. M. & Weinstein, C. S. (Eds.), *Handbook of classroom management: Research, practice, and contemporary issues.* Mahway, NJ: Lawrence Erlbaum Associates.

StataCorp (2011). *Stata Statistical Software: Release 12*. College Station, TX: StataCorp LP.

Sughrue, J. A. (2003). Zero tolerance for children: Two wrongs do not make a right. *Educational Administration Quarterly, 39*(2), 238-258.

Townsend, B. L. (2000). The disproportionate discipline of African American learners: Reducing school suspensions and expulsions. *Exceptional Children, 66*(3), 381-391.

U.S. Department of Education, National Center for Education Statistics. (1998) *Violence and Discipline Problems in U.S. Public Schools: 1996-97*, NCES 98-030, by Sheila Heaviside, Cassandra Rowand, Catrina Williams, and Elizabeth Farris. Project Officers, Shelley Burns and Edith McArthur. Washington, DC. Retrieved from: http://nces.ed.gov/pubs98/98030.pdf

Zehr, M. A. (2010). *Education Week.* Obama administration targets ‘disparate impact' of discipline. Retrieved from: http://www.edweek.org/ew/articles/2010/10/07/07disparate_ep.h30.html

Improving the Estimation of the Risk of School Suspension Using Continuous-Time Survival Analysis: A Case Study in the Public Middle Schools of One Metropolitan Region

Stephen L. Hoffman

Harvard Graduate School of Education

Improving the Estimation of the Risk of School Suspension

Using Continuous-Time Survival Analysis:

A Case Study in the Public Middle Schools of One Metropolitan Region

For 10 years I was a middle-school principal and a high-school assistant principal. On most days of the school year, I assigned formal discipline to the handful of students whose misbehavior was brought to my attention. Many times, I suspended students out-of-school, following district procedures and the approach to discipline advocated by my superiors. Furthermore, since my first day as an administrator, I was made acutely aware of the racial dynamics in schools, particularly regarding the disproportionality in school discipline for Black and Hispanic students that has also been documented extensively in recent scholarship (Beck & Muschkin, 2012; Gregory, Cornell, & Fan, 2011; Kinsler, 2011; Petras, Masyn, Buckley, Ialongo, & Kellam, 2011; Skiba et al., 2011).

My experiences as a school principal and my subsequent reading of the scholarly literature have motivated me to scrutinize disciplinary policies in schools more formally. Thus, here in my qualifying paper (QP), I have conducted a case study to explore the out-of-school suspension of students in the public schools in one metropolitan region. In particular, I bring an innovative analytic framework—survival analysis—to bear, to improve the estimation of suspension rates and describe the timing of suspension using a finer lens than others have done. Broadly speaking, I extend prior work on school discipline by investigating *whether* middle-school students are suspended from school, and if so, *when, during the school year,* these school suspensions occurred.

I begin by analyzing when, during the course of the 2009-10 academic year, students were at the greatest risk of being first suspended from school (that is, given that

they had not been suspended at an earlier point in the academic year). I explain my use of the term risk more formally, both in terms of *who* is at risk, and also *when* students are at relatively higher or lower rates of occurrence of first suspension. Given that I have worked with data in which time to suspension was measured continuously, I fitted a series of Cox-regression models (Cox, 1972) to first investigate whether there were differences in this risk, by student grade level. Then, I document the continued pattern of disproportionality in the risk of school suspensions by race.

For the last 40 years, researchers in fields like criminology (e.g. Dugan, Lafree, & Piquero, 2005) medicine (e.g. Heikinheimo et al., 2013), and sociology (e.g. Zheng & Thomas, 2013) have used the strategy of Cox (1972) regression analysis to study analogous topics. Cox-regression analysis is certainly the most popular technique for modeling rates of event occurrence in continuous-time, because it is both elegant and computationally practical (Cleves, Gould, Gutierrez, & Stata Corporation., 2010). In fact, as of March 7, 2014, Cox's seminal paper detailing this method had been cited 27,942 times on the Web of Science! However, I am unaware of any other researcher who has applied Cox-regression analysis to the investigation of the phenomenon of out-of-school suspensions, as I have here.

I have organized the QP into five major sections, following this introduction. First, in the *Background and Context* section, I review the literature to provide a rationale for my investigation of *whether* students have been suspended, historically, from school. I focus on disparities by race and ethnicity generally, and for middle-school students in particular. In this section, I argue that survival analysis is a plausible strategy for addressing the "when" question, in estimating the risk of suspension during the school

year. I end this section by stating my specific research questions. Second, in my *Research Design* section, I describe the research site, the features of the dataset, the sample of students, and the procedures that I use to address my specific research questions. Third, I present my *Findings* from this analysis. Fourth, I discuss the limitations of my analysis in a section on *Threats to Validity.* Fifth, I synthesize and review the meaning and significance of the findings in a *Discussion* section.

Background and Context

Out-of-School Suspension as a Disciplinary Tool

The use of out-of-school suspension of students as a disciplinary tool has been a common practice by public-school administrators for decades (Edelman, Beck, & Smith, 1975; Wu, Pink, Crain, & Moles, 1982; Arum, 2003). However, prior to the 1970s, the use of school exclusion—such as out-of school suspension—was almost entirely informal and discretionary. Typically, school principals were not subject to formal oversight of their use of suspensions, even by school superintendents (Edelman et al., 1975). Then, during the 1960s and 1970s, the issue of school suspensions came to the forefront of public debate, as a result, in part, of the discussion of racial issues spearheaded by the Civil-Rights movement (Arum, 2003). This led to substantial new policy decisions about the use of school suspension, ordered by the courts and sanctioned by the federal government (Edelman et al., 1975; Ornstein, 1982). In the *Goss v. Lopez* (1975) decision, the U.S. Supreme Court settled whether students might be suspended out-of-school, due to their behavior.

Goss v. Lopez. The *Goss v. Lopez* (1975) decision remains a signature legal holding about the use of out-of-school suspensions (Arum & Preiss, 2009). In this ruling,

the U.S. Supreme Court described how school staff should enforce standards of conduct for students because some "modicum of discipline and order is essential if the educational function is to be performed" (*Goss v. Lopez*, p. 580). Noting that disciplinary events in schools were frequent occurrences, requiring "immediate, effective action," the court explained that the use of out-of-school suspension was considered a necessary tool to maintain order (*Goss v. Lopez*, p. 580). By reacting to the social and political issues of the times, the Supreme Court outlined a broad framework for the use of school exclusion that still guides most public-school policies about out-of-school suspensions (Arum & Preiss, 2009).

But in addition to sanctioning the use of out-of-school suspension—the rights of school officials— the Supreme Court also outlined rights for students. Thus, the *Goss v. Lopez* (1975) ruling affirmed that students also possess rudimentary due-process rights regarding school exclusion. Even when a student is suspended out-of-school for violating school rules, school administrators must not act arbitrarily or capriciously when doling out the punishment. As Justice White explained in the ruling, students who are suspended from school "have interests qualifying for protection of the Due-Process Clause" (p. 581). Consequently, school officials, upon suspending a student for 10 days or less following a disciplinary incident, must explain explicitly *why* a student is to be suspended. School staff must specify what evidence the school used to make the decision to suspend. And, students must be afforded an opportunity to tell their side of the story.

Federal Responsibilities/States' Responsibilities. Despite political arguments about the federal role in public education in the United States—e.g. recent debates about No Child Left Behind (2002), Race to the Top (2011), or the implementation of the

Common Core Standards (2010)—education remains primarily a function and responsibility of the states. All 50 states in the U.S. maintain a public-school system. Even after the turmoil in the wake of the Brown v. Board of Education (1954) ruling ordering school desegregation, state governments did not seriously question whether states would continue to support public education (Butts, 1955). While states like Indiana—an early adopter of the *Common Core Standards*—are now withdrawing their support for implementing them (e.g. Martin, 2014, reporting for the New York Times) the broader framework of federal and state responsibilities remains.

State-government policies stipulate the rules and regulations regarding school discipline, including policies about the use of out-of-school suspensions. In *Goss v. Lopez* (1975) the Court ruled specifically that the state of Ohio (as this case dealt specifically with a difficult disturbance at a high school in the Columbus, Ohio, public-school system) is "not constitutionally obligated to establish and maintain a public school system" (p. 574). But since each has chosen to extend this right of a public education, states must follow the Supreme Court's broad guidelines concerning due process when using out-of-school suspension as a disciplinary tool.

Thus, any study of school discipline broadly—and of the use of out-of-school suspensions in particular—must consider the context present in the home state of the schools. For example, in the current study, I investigated the application of student suspension in a non-random sample of schools, in one particular state. The relevant state law regarding the use of out-of-school suspension asserts that school officials may suspend a pupil for not more than five school days, unless a notice-of-expulsion hearing has been sent. Thus, principals, charged with maintaining order and discipline, *may* use

out-of-school suspension for students violating school rules. However, questions about w*hether* students may be suspended for violating school rules are left frequently to the judgments of school principals—a situation similar to schools described by Edelman, et al (1975) 40 years ago.

***Whether* students were suspended from school?** As Justice White maintained, out-of-school suspension was "considered not only to be a necessary tool to maintain order, but a valuable educational device" (*Goss v. Lopez*, p. 580). Consequently, large numbers of students are suspended from the public schools in the U.S. every year, although summaries of the rates at which these suspensions are administered depend on the methodology used to estimate them. For example, the federal government authorizes the mandatory collection and summarization of data on school suspensions through the United States Department Office of Civil Rights (2013). These guidelines define out-of-school suspension as "excluding a student from school for disciplinary reasons for one school day or longer" (p. 35). A recent summary provided by the National Center for Education Statistics (NCES) estimated that 3,325,000 students were suspended out-of-school at least once during 2006 (U.S. Department of Education, 2012). Thus, about 7% of all public elementary- and secondary-school students were suspended in that year. And the NCES summary noted that students were counted only once in the estimates, even if students were suspended multiple times from the same school during the same school year. Furthermore, these NCES estimates of suspension rates differ considerably by state. In North Dakota, at one extreme, approximately 2.2% K-12 public-school students of were suspended out-of-school in 2006. At the other extreme, in South Carolina, the percentage of students suspended at least once was 11.9%.

***Why* were students suspended from school?** In addition to the question of *whether* students are subject to out-of-school suspensions, formal school discipline is administered for a large range of kinds of disobedience. *Why* a student was suspended from school—for events that are moderately disruptive, like insubordination or disruption; to more serious infractions, such as fighting or using drugs or alcohol—has been the subject of scholarly interest for decades (Theriot & Dupper, 2010; Raffaele Mendez & Knoff, 2003; Skiba, Peterson, & Williams, 1997; Wu et al., 1982; Edelman et al., 1975). While in the current paper, I focus on questions of *whether* and *when* students are suspended, the reasons *why* a student was suspended from school will also become an important aspect of my future work, particularly when questions about disproportionality in school discipline by race come to the fore.

In the wake of the Civil-Rights movement, the *Office of Civil Rights* first began collecting data about the nature and use of school suspension in the United States (Edelman et al., 1975). Thus, immediately following the *Goss v. Lopez* (1975) ruling, cited above, advocates like those at the *Children's Defense Fund* attempted to make sense of the myriad reasons that students were being subject to out-of-school suspensions. For instance, in their report, Edelman, Beck and Smith (1975) found that most suspensions "were for nondangerous offenses," such as "insulting" the teacher or tardiness to class—and they recommended that schools reduce the use of out-of-school suspension for these behaviors (Edelman, Beck & Smith, 1975, pp. 37-38). They concluded that patterns of suspension were largely a consequence of differences among school administrators and suspension policies, rather than different behavioral patterns among students. This was confirmed empirically by Wu, et al (1982) in a subsequent national study of 641 public

secondary schools conducted in 1976. Wu and his colleagues concluded that student misbehavior was only one factor resulting in an out-of-school suspension. Principals' decisions to suspend students were mediated by teachers' perceptions and beliefs, schools' administrative structures, and institutional biases (Wu et al., 1982).

Several recent studies have also described the types of infractions that have led to out-of-school suspension for students. In a study of disciplinary interventions in middle schools, for instance, Skiba, Peterson, and Williams (1997) noted that the reasons why students were suspended out-of-school differed considerably among principals. While fighting seemed to be a straightforward case for the use of out-of-school suspension, rates of suspension for insubordination or disruption differed considerably by school. In another study, Raffaele Mendez and Knoff (2003) also studied out-of-school suspensions in a large school district in Florida and found that about four-times as many suspensions were handed out for "disobedience," than for "violence against persons." Very few suspensions were administered for the possession and use of alcohol, drugs, or weapons—particularly in middle schools. Similarly, Theriot and Dupper (2010) studied discipline problems during the transition from elementary to middle school in a medium-sized school district. They found that less than half of the out-of-school suspensions were due to infractions like fighting, threatening behavior, or theft. More commonly, out-of-school suspensions were imposed for reasons like "conduct prejudicial to good order" (p. 213). Thus, while issues of drugs, weapons, and gang activity presented serious issues that school principals must address, disruptive and insubordinate conduct were behaviors that were addressed as often as daily.

At what age are students suspended from school? Adolescence can be a particularly trying time for children, as well as for those who teach or otherwise staff a middle school. Few studies have investigated how rates of out-of-school suspension differ by grade level. Studies have indicated that suspension rates for middle-school students exceed those of elementary-school students substantially; and, in some studies, rates of suspension for middle-school students have even exceeded rates for high-school students (Theriot & Dupper, 2010; Losen & Skiba, 2010; Raffaele Mendez & Knoff, 2003). Losen and Skiba (2010) explained that, as of 2010, no nationally representative reports based on federally collected discipline data have been disaggregated by grade level. However, Losen and Skiba (2010) reported—in their study of suspensions in 18 large urban school districts—that the average suspension rate in middle schools in these districts was 11.2%, in 2006. Similarly, in their study of students in a public-school district in Florida, Raffaele Mendez and Knoff (2003) reported that the percentage of students who experienced at least one suspension during the 1996-97 year was less than 4% for elementary-school students, 24% for middle-school students, and 18% for high-school students. And, as Theriot and Dupper (2010) noted in their study of the transition from elementary school to middle school, rates of suspension for students in 6th grade (at the beginning of middle school) were substantially higher than the rates of suspension for the same students during the previous school year (when the students were still in elementary school).

Refining Annual Suspension Rates: *When* Are Students Suspended During the School Year?

In his study of school principals, Lortie (2009) noted that the day-to-day work of principals is marked by both urgency and randomness. "Outbursts of student misbehavior can be sudden and unpredictable" (Lortie, 2009, p. 123). Thus, one might hypothesize that suspensions must be distributed essentially at random over the school year, and be administered uniformly. However, typically, the question of *when* suspensions occurred—that is, at what point during the school year—has been left unanswered in investigations of school suspensions. Are 6th-grade students at greatest risk of experiencing suspensions at the start of the school year, when, perhaps, 6th-grade students must acclimate immediately to the rigors of middle school? Or, are the first suspensions of middle-school students more evenly spaced out during the course of the year? Do 8th-grade students tend to become unruly towards the end of the academic year, experiencing substantial increases in the risk of suspension? In their concluding remarks, for instance, Theriot and Dupper (2010) speculated that the rate of suspensions of 6th-grade students would decrease after students had time to adjust to middle school, although this hypothesis remained unaddressed in their research.

While the estimation of simple annual rates of suspension, as carried out in many studies, provided some information about *whether,* and *why,* students are suspended from school (Raffaele Mendez & Knoff, 2003; Theriot & Dupper, 2010), information about *when* during the school year suspensions have occurred is notably absent from most research. One innovative approach to investigating *when* students are suspended from school is by the use of the methods of survival analysis. Although indirectly, survival

analysis focuses on, and summarizes, the *time* that elapses from an "origin" (such as the beginning of a school year) until the occurrence of a "failure" (such as an event like a school suspension). Thus, using survival analysis, I can examine both *whether*, but also, *when*—during the school year—a student was at greatest risk of being suspended from school (Singer & Willett, 2003). Modeling the day-by-day risk of school suspension allows me to determine when the occurrence of school suspension is at its greatest (conditional) risk of occurring (Singer & Willett, 2003).

As a result, I am able to answer interesting questions about school suspension. For example, I can ask: Do suspensions occur earlier in the school year, tending to set boundaries on student behaviors, as Theriot and Dupper (2010) hypothesized for 6th-grade students? Or, are suspensions more akin to random incidences that occur uniformly throughout the school year, with students socialized into complying with school norms more gradually, as Lortie (2009) explained? Losen and Skiba (2010) described rule-breaking as normal behavior—particularly for middle-school students, who are accustomed to challenging authority both at home and at school, as part of their normal adolescent social development. So, while some behavioral norms may be established early in the school year, perhaps other developmental factors continue to play a substantial role throughout middle school.

Does survival analysis describe all incidents of event occurrence? Many studies, such as the NCES report (U.S. Department of Education, 2012), cited earlier, state that students were counted only once in the estimates, even if students were suspended multiple times from the same school during the same school year. However, in one recent exception, Sullivan et al (2013) reported that the proportion of students who

were suspended *exactly* once during the school year was slightly larger than the proportion of students who were suspended multiple times.

One strategy to overcome this blurring of single and multiple suspensions is to focus on the *first* suspension of the school year. For example, Petras, Masyn, Buckley, Ialongo, and Kellam (2011), in a study of who is most at risk of school removal, focused on the *first* school suspension (in a student's school career), because they hypothesized that the timing of the first suspension was predictive of later school problems. Thus, they used discrete-time survival analysis to investigate the risk of suspension from school *for the first time*. However, Petras and colleagues focused on *whether* and *in which grade* a student was first suspended, during an entire student's elementary-school career (estimating a *yearly* probability of out-of-school suspension from grades 1 through 7). In contrast, I focus on the *daily* risk of students being suspended for the first time, out-of-school, during the course of one school year, for three cohorts of students in their 6th, 7th, and 8th grades, respectively.

The problem of censoring. One of the critical problems facing scholars who wish to investigate the risk of out-of-school suspension is the ubiquitous problem of *censoring*. In this study, in which data on school suspensions were collected and analyzed for three cohorts (6th, 7th, and 8th grade) of students over one school year, the students fall into two categories. The first category includes those who were suspended for the first time while they were observed during the school year. They provide direct insight into the question of *"whether"*, and the answer for them is "yes." The rest of the students in the three cohorts fall into the category of *censored observations* (Singer & Willett, 2003). For these latter students, I do not observe a first suspension during the school year. The

challenge that I face in estimating the risk of suspension is to incorporate both uncensored and censored students legitimately into the same analysis in order to obtain unbiased estimates of suspension rates, by group.

The *risk* of first suspension. Assuming that all middle-school students are at eligible to be suspended from school, using survival analysis allows me to investigate when*,* during the school year*,* the risk of school suspension is the highest. In such research, the fundamental quantity of interest is called the *hazard*—the risk at any time during the school year that a student might be suspended, given that the student has not yet experienced a first suspension. Furthermore, in this study, where I have chosen to model this phenomenon in continuous time (because the student records are available to the nearest day), I estimate and report the *hazard rate*—the conditional probability of first suspension, *per day* in school (Singer & Willett, 2003).

Differences in Suspension From School By Race or Ethnicity

Historically, public schools in the United States had served, primarily, the needs of White students, both male and female. This was particularly the case for students enrolled in public education beyond primary school. Indeed, prior to the Civil War, more than 99% of students aged 12 years old in the nation's public schools were White students.[8] As Goldin and Katz (2008) explained, free public schooling has been provided for White youth of both genders and all social classes since the late 1800s. Average levels of educational attainment rose throughout the first two-thirds of the 20th century, with the

[8] Author's estimation using U.S. Census data, archived by the Minnesota Population Center at the University of Minnesota (Ruggles et al., 2010).

proportions of males and females attaining a high-school diploma in school rising throughout that time at roughly comparable rates.

While public schools served the vast majority of Whites in the U.S., educational opportunities for non-Whites were considerably different, traditionally. For example, Anderson (1988) noted that for at least 70 years following the Civil War, a substantial portion of the White population resisted efforts to provide even basic education to Black youth, or endorsed education primarily for preparing Blacks to work in a perpetually lower-class status. Similarly, Lomawaima (1999) described a long, sordid history of providing educational "opportunities" for Native-American children—including a 450-year history of boarding schools and other American institutions teaching a "special pedagogy" formulated by Whites. Thus, for both the historically enslaved Black population and the indigenous Native-American population, public educational opportunities lagged considerably behind those for Whites.

The presence of the two other major racial/ethnic minorities, Asian and Hispanic, was quite small until recent demographic shifts. Tamura (2001) noted that Asian Americans have lived in the U.S. for over 150 years. But the proportion of Asians Americans in the U.S. population remained less than 1% before the end of World War II. And beginning in 1970, the U.S. Census first separated out questions about "race" from questions about whether a person was Hispanic, or not (U.S. Census, 2002). This disaggregation of "Spanish/Hispanic origin or descent" (1980 wording) from country of origin is also mandated by the U.S. Office of Management and Budget (1997), and is now used in the recent decennial censuses and the American Community Survey.

However, the proportions of these two demographic groups in the population have grown substantially in recent years. Between 1980 and 2005, the public-school population of Asians/Pacific Islanders (hereafter, "Asian") has grown by 260 percent (KewalRamani, Gilbertson, Fox, Provasnik; 2007). And, the part of the U.S. population of Hispanic ethnicity is now the largest minority population—surpassing the Black population—as of the turn of the 21st century (KewalRamani et al., 2007). Thus, in summary, as of 2012, the percentage of White public middle-school students in the U.S. was 55%, while 5% were Asian, 16% Black, and 23% Hispanic.[9]

Racial and ethnic identities and the differences among them are unstable concepts, constantly under dispute and transformation (Omi & Winant, 1994). During the first half of the 1800s, all people were counted by age, sex, and "color" (the choices were White, Black, or Mulatto) in the U.S. census (U.S. Census, 2002). And, as explained above, the modern conception of "Hispanic" was formulated in the middle of the 20th century as an ethnicity, rather than a race.

Similarly, since 1977 the federal government has also collected data systematically about the ethnicity and race of students. And, since 2007, these federal guidelines specifically ask students (or their parents) a two-part question about students. First the respondent (a parent or guardian, or the child herself/himself) is asked to identify the student's ethnicity as either Hispanic or not. Second, the respondent is asked to identify his or her race (White, Black or African American, Asian, American Indian or

[9] Author's estimation using U.S. Census data, archived by the Minnesota Population Center at the University of Minnesota (Ruggles et al., 2010).

Alaska Native, Native Hawaiian or Other Pacific Islander) or races (U.S. Department of Education, 2008).

During the last four decades, the demographic composition of public schools has changed substantially. In 1970, approximately 82% of all public-school middle-school students were White. About 13% were Black students, with less than 4% Hispanic. Asian and Native-American students comprised less than 1% of public-school middle-school students. However, 40 years later, in 2010, approximately 55% of all public-school middle-school students were White. About 16% were Black, with more than 23% Hispanic, and nearly 5% Asian.

Racial/ethnic disparities in school suspension. Decades of research have also documented important racial/ethnic disparities in the rate at which students are suspended out-of-school (Edelman et al., 1975; McCarthy & Hoge, 1987; Skiba et al., 1997; Townsend, 2000; Losen & Skiba, 2010; Kinsler, 2011; Petras et al., 2011; Skiba et al., 2011; Sullivan et al., 2013). As Skiba et al (2011) commented, in a study of Black and Hispanic disproportionality in school discipline, "race is not neutral." By analyzing school disciplinary records in 364 elementary and middle schools, the authors, using logistic regression analysis, concluded that the probability that either Black or Hispanic students would be suspended from school was greater than for White students, for the same or similar misbehavior. Skiba and his colleagues also point out, in their study, that they drew a sample of schools that implemented one particular data-collection system, but they provided evidence that the sample represented the U.S. population of public-school students along dimensions of gender, race/ethnicity, and special education.

In another study also focused primarily on disproportionality in school suspension by race/ethnicity, Wallace, Goodkind, Wallace, and Bachman (2008) analyzed data collected from a nationally representative survey of high-school students. Using logistic regression analysis, they estimated that the odds that Black students reported being suspended from school were more than three times the odds for White students. Comparing Native-American students (a category not explored by Skiba et al) to White students, the odds were twice the odds for Whites, while the odds-ratio comparing Hispanic students to White students was 1.7. Wallace et al (2008) and other researchers (Morris, 2005; Sullivan et al., 2013) have found that Asian students, as a demographic group, have lower odds for reporting suspension, compared to White students. While these two studies differ in data-collection methodology—with Skiba et al (2011) utilizing administrative records, and Wallace et al (2008) using individual survey data—both studies highlight the continued pervasiveness of disproportionate risk of school suspension by student race. This disproportionality has changed little since the research of Edelman et al (1975) four decades ago. And so, in the current research, I have focused particularly on differences in the risk of school suspension by race/ethnicity.

But, why are Black students, as a demographic group, at the greatest risk of being suspended (Skiba et al., 1997; Townsend, 2000; Raffaele Mendez & Knoff, 2003; Nichols, 2004; Hinojosa, 2008; Kinsler, 2011; Sullivan et al., 2013)? Kinsler (2011) attributed much of the differences in suspension rates between Black and White students to differences in the schools they attend. Most other researchers (Ferguson, 2000; Gregory et al., 2011; Hinojosa, 2008; Payne & Welch, 2010; Skiba et al., 2011; Sullivan et al., 2013) concluded that differences in the rate of suspension, by race, were attributed

to pervasive differences in the school experiences of Black and White students. For example, Gregory, Cornell, and Fan (2011) combined their analysis of Virginia state records of school suspensions with the results from a survey of 9th grade students, sampled from more than 90% of the high schools in Virginia. They concluded that suspension rates for Black students and White students, which differ substantially, were not attributable to differences in either school size or SES. Rather, their findings are more consistent with those of Ferguson (2000) and Payne and Welsh (2010)—that the school experience for Black students is decidedly different, with Black students being subject to more punitive discipline than other students who committed the same offenses.

Less research has been conducted concerning the punishment of Hispanic students in public schools, in the United States.[10] However, several studies have also found that the rate of suspension for Hispanic students is higher than that for White students. For example, in addition to the differences in the rates of suspension they found between Black students and White students, Skiba et al (2011) also found that the odds of suspension for middle-school Hispanic students were higher than for White students. In another study of middle-school suspensions, conducted in the primarily Hispanic and Black schools in Miami, Florida, Arcia (2007) found that rates of suspension for Black students were higher than for Hispanic students. Finally, Sullivan. Klingbeil, and Van Norman (2013) concluded that the odds of school suspension for Hispanic students, while lower than for Black students, was higher than the odds of school suspension for White students.

---

[10] For the results in this paper, I use the term Hispanic, following the common demographic shorthand descriptor used by many. Other researchers prefer the term Latino/a.

Conclusion: Specific Research Questions

In the above discussion of the background and context of my research, my predominant focus has been on studies that report estimates of the probability ("whether") a student will be suspended in a given academic year—known as the "suspension rate." The NCES (2012) report, described above, for example, estimates explicitly the probability that students were suspended out-of-school at least one time ("whether"). And, several studies have noted high rates of suspension for middle-school students, as compared to other public-school levels (Theriot & Dupper, 2010; Losen & Skiba, 2010; Raffaele Mendez & Knoff, 2003). Consequently, the results of these studies, as well as my own professional experience as a middle-school principal, incentivize me to focus on similar probabilities—that is, *whether* middle-school students have been suspended from school, in this research.

However, as Lortie (2009) remarked, students are "only gradually socialized into complying with the norms of orderly behavior" (p. 123). Therefore, I believe that it is also important also to analyze *when*, during the school year, these suspensions occur. For school personnel faced with the ever-changing landscape of how the experiences of students in school change over time, understanding who is subject to school suspension, and *when* during the school year these occur, seems to me to be fundamental. Furthermore, I hypothesize that the risk of suspension likely differs by student race/ethnicity, and thereby potentially influences decisions made by school principals about whether students *should* be suspended from school.

Therefore, in this research, I conduct a case-study in the public middle schools of one metropolitan region to describe *whether,* and if so, *when*, middle-school students

were first suspended during the course of a single school year, and how the occurrence of first suspension differed by student race/ethnicity. My specific research questions are:

1. How does the risk of first suspension differ for middle-school students, over the course of a school year? In particular is the risk of suspension higher at the beginning of the school year than it is at the end?
2. Is the annual temporal profile of risk of first suspension higher, and of a different shape, for Black and Hispanic students than for White students?

I address these questions by capitalizing on a rich and extensive database provided to me by a state agency. As my database contains the daily history of suspensions over one academic year, for three cohorts of students (beginning the school year at the 6th, 7th, and 8th grade levels), when I address each research question, I also investigate implicitly whether the within-year profile of risk of first suspension differs by grade level.

Research Design

Site

For this case study, I focus on students enrolled in all public middle schools located in one metropolitan region surrounding a medium-sized city. I have concealed the name of this city to protect the confidentiality of the students whose middle-school careers are examined. I argue that this was an ideal site for my case study for several reasons. First, the sample was large enough to provide a sizeable number of incidents of out-of-school suspensions—a subsample of students comparable in size to other research on this topic, reviewed above (e.g. Raffaele Mendez & Knoff, 2003; Theoriot & Dupper, 2010; Sullivan et al, 2013). Second, I have access to information about the students' grade-level in school, allowing me to compare risk of first suspension among students in

the 6th, 7th, and 8th grades, over one school year, in middle school. Third, my sample contains racial/ethnic diversity ample enough to support the investigation of any disproportionate risk of suspension for Black and Hispanic students, compared to White students.

This metropolitan region contained approximately a half-million residents. Approximately four-fifths of the residents were non-Hispanic White, while Asian, Black and Hispanic residents constituted roughly equal proportions of the population. The largest district served more than 20,000 students from kindergarten through 12th grade. Each of the other districts was smaller and served less than 6,000 students. Most of the school districts served 6th through 8th grade students in a traditional middle-school configuration, but a few of the districts served only 7th and 8th grade students in a two-year configuration of middle schools. Consequently, I excluded approximately 350 6th-grade students who attended "intermediate" schools in these districts.

Dataset

I have been provided confidential access to the educational records of all of the middle-school students enrolled in any middle school in this region at the beginning of the 2009-10 school year. My dataset contains information on the date each student was enrolled in school, was first suspended out-of-school (if suspended at all), left school for another reason (perhaps moving to another school), or completed the school year without formal discipline by school officials. I constructed this merged dataset by combining information from three linked datasets containing individual student records that I obtained, with permission, from a government agency.

The first dataset that I integrated into my larger data assembly consisted of student-enrollment information. It contained the date of enrollment and departure from each of the public middle schools in the region, for each student, identified by a unique student number. More than 15,000 students attended one of the middle schools in the region for at least a portion of the school year. Students enrolled and left schools throughout their school career. However, in this sample, 97% of the students who were enrolled in one of these middle schools during the first three weeks of school remained enrolled until the end of the school year. Consequently, in my analysis, I focus on the records of the 13,256 students who enrolled at one of the schools within the first three weeks of the school year.

The second dataset that I incorporated into my work contained demographic information. The contents of this dataset provide information on the two variables that I treat as time-invariant covariates in my analysis. In this dataset, student grade-level was recorded as each student's grade at the end of the school year. Additionally, schools recorded students' race/ethnicity when students were enrolled initially into the public school, and students—or their parents—identified their primary racial/ethnic classification.

The third dataset that I incorporated into my analysis recorded the date of each disciplinary incident that was reported to the state by the schools and districts in this sample, by student ID. Thus, it contains information about in-school suspensions, out-of-school suspensions, and expulsions reported by the schools and districts. But, for the current investigation, I limit my investigation to the *first* occurrence of out-of-school suspension of the school year, for any student. I do this—in a similar vein to Petras,

Masyn, Buckley, Ialongo, and Kellam (2011)—because the timing of suspension may predict subsequent school disciplinary issues. In my subsequent research, in my doctoral thesis, I plan to investigate the occurrence and timing of school suspension longitudinally over each student's entire middle-school career, incorporating information on the first and all subsequent suspensions in a multiple-spell survival analysis.

Sample

I analyze data on an analytic sample containing 13,256 middle-school students, who attended the 31 schools at my site, over the 2009-10 school year. Following Kinsler (2011), I included students attending all public middle schools that served students in grades 6th through 8th, or grades 7th and 8th, exclusively.

Measures

As is required for continuous-time survival analysis, I formatted my combined dataset as a person-level dataset (Singer & Willett, 2003) with each sampled student contributing one row of data. In each row, in addition to ID codes that identified the individual student, school, and district, I coded the values of the following variables:

Outcome:

Conceptually, in Cox-regression analysis, the analytic outcome is an expression of the risk of suspension, called the hazard rate. It is defined as the conditional probability that an individual will experience the event of interest—in this case, a first suspension from school—per unit of time (Singer & Willett, 2003). It is constructed implicitly during the analysis from the values of the following two outcome variables:

TIME (*t*) is a continuous variable that records the passage of time, measured in school days, from the origin (the student's entry date into the school), until a student

experienced a first suspension from school during the 2009-10 school year, or was censored. Note that, in my sample, 92.3% of the students were *censored*—that is, they did not experience a first suspension while observed—either because they moved away on a specific day during the school year before any record of suspension, or because they completed the school year without being suspended from school at all.

FIRSTSUS is a dichotomous variable, coded as 1 if a student experienced a first suspension from school during that year, and 0 if not suspended (the latter being the *censored* students).

Question Predictors:

GRADE is a vector of three time-invariant dichotomous predictors (GRADE6, GRADE7, GRADE8) to record the grade in which each student was enrolled during the year of the investigation. Each variable is coded 1, if the student was enrolled in the respective grade, 0 otherwise. I omit the 6^{th} grade category from my statistical models to provide a reference category.

RACE is a vector of time-invariant dichotomous variables that record the primary race/ethnicity of each student, using definitions provided by the state. The vector includes the variables ASIAN, BLACK, HISPANIC, NATIVE, and WHITE, each coded 1 or 0, to identify the student's race/ethnicity. I omit the predictor WHITE from my statistical models to define a reference category.

Data-Analytic Plan

In my dataset, the occurrence of out-of-school suspensions was recorded by the state to the nearest day, guiding me to use continuous-time survival analysis to address

my research questions. While parametric approaches to continuous-time survival analysis (e.g. Weibull, 1951) might also shed light on the issue, for this paper, I elected to apply a popular semi-parametric method—Cox-regression analysis (Cox, 1972)—as a strategy for describing differences in the risk of suspension among students from different groups. The Cox-regression strategy is a popular technique but may seem to be a puzzling analytic choice for this study, because it assumes no particular parameterization of the relationship between hazard rate and time. In fact, if the baseline hazard rate were left un-estimated, Cox-regression would provide no information about *when* the risk is highest. However, once I include the two question predictors—*GRADE* and *RACE*—into the model, comparing the estimated of risk of first suspension between groups provides a sensible method to analyze the phenomenon. In addition, the temporal dependence of the hazard rate on time can be estimated post-hoc in the data using conventional methods of Kaplan-Meier (1958) estimation. The obtained sample risk profile—combined with the findings of grade and race-dependence from fitted Cox-regression models—provide answers to my research questions.

In each Cox-regression model that I specify and fit below, my outcome was the (log) hazard rate. The right-hand side of each hypothesized model then contained an unspecified (log) baseline hazard function (that was not estimated directly during the model fitting, but was recovered from the sample data post-hoc), plus a weighted linear combination of the effects of the hypothesized predictor variables, each accompanied by its respective slope parameter. In each model, I accounted for the clustering of students in schools by adding a random effect of school, or "frailty," (Vaupel, Manton, & Stallard, 1979) that acknowledges differences in estimated risk attributable to what school a

student attended. Finally, I used the Efron (1977) method for handling tied outcomes, when two or more students have identical estimated times to the event.

RQ1: How does the risk of first suspension differ for middle-school students, over the course of a school year? In particular, is the risk of suspension higher at the beginning of the school year than at the end?

As my sample of public-school students contains middle schools with students at three grade-levels, I first began by investigating whether the within-year profile of risk of first suspension differed *in level* among 6th, 7th, and 8th grade students. To describe the distribution of the risk of first suspension over the school year, I fitted the following Model A, in the full sample:

$$\ln h\left(t_{ijk}\right) = \ln h_0\left(t_{ij}\right) + \boldsymbol{\alpha}'\boldsymbol{GRADE}_{ijk} + \upsilon_k \qquad (1)$$

where $h\left(t_{ijk}\right)$ represents the hazard rate describing the risk of first suspension of student *i*, on day *j*, in school *k*. Parameter $\upsilon_k$represents the random effect of school *k*, and was assumed to be gamma-distributed.

Parameter vector $\boldsymbol{\alpha}'$represents the population fixed effects of student grade on the log-hazard-rate of first suspension, thereby permitting the *level* of first-suspension risk to differ by grade. If an estimate of one of the elements of this vector is positive and statistically significant, then I know that the hazard rate describing first suspension is elevated in the corresponding grade. I have treated 6th grade as a reference category. Thus, if an estimate of α_1 was positive and statistically significant, then I concluded that the risk of first suspension was higher for 7th grade students than for 6th grade students.

Note that, in this model and in the standard Cox-regression analysis, the effect of *GRADE* is specified as a main effect only. Then, anti-logging the associated parameter

estimates, the several fitted hazard rate profiles in each grade, defined by parameter vector $\boldsymbol{\alpha}'$ have similar shapes and are *proportional* to each other. To facilitate interpretation of these latter model parameters, in my tables of fitted models, I have also included a column containing the anti-logged parameter estimates that can then be interpreted as *hazard ratios.* Each hazard ratio (the anti-log of α_1, say, corresponding to 7th grade) multiplies the *risk* of first suspension in the reference 6th grade to obtain the *risk* of first suspension for 7th grade students. A similar interpretation comparing 8th graders to 6th graders applies for α_2.

Finally, to this baseline proportional-hazards model, I tested the addition of an interaction between *GRADE* and linear *TIME,* to investigate whether the proportional-hazards assumption was violated and therefore that any pattern of yearly risk differed by grade, in *shape* as well as *level*. After exploring various parameterizations, I found that the following non-proportional hazards model that included an interaction between each *GRADE* and linear time, was the most appropriate to fit the data:

$$\ln h\left(t_{ijk}\right) = \ln h_0\left(t_{ij}\right) + \boldsymbol{\alpha}'\boldsymbol{GRADE}_{ijk} + \boldsymbol{\gamma}'(\boldsymbol{GRADE}_{ijk} \times TIME) + \upsilon_k \qquad (2)$$

and I obtained the corresponding estimated parameters and fit statistics.

In the semi-parametric Cox approach to continuous-time survival analysis, the baseline hazard rate profile, contained as an "intercept" in the model, requires no particular parameterization and is not estimated directly during model fitting. However, after model fitting, it can be recovered from the data, using the Kaplan-Meier (1958) estimator, and plots of the fitted risk over time can be displayed, appropriately shifted according to the magnitudes of the estimated elements of parameter vector $\boldsymbol{\alpha}'$. Thus, after fitting this non-proportional hazards model, I was able to obtain and present plots of the

fitted risk of the *TIME* to first suspension, by *GRADE,* throughout the school year, to summarize the occurrence and timing of first suspension.

RQ2: Is the annual temporal profile of risk of first suspension higher, and of a different shape, for Black and Hispanic students than for White students?

To address my second research question, I added my principal question predictor—describing the annual temporal profile of risk for students of different racial/ethnic categories—to the model fitted above, as follows:

$$\ln h(t_{ijk}) = \ln h_0(t_{ij}) + \boldsymbol{\alpha}'\boldsymbol{GRADE}_{ijk} + \boldsymbol{\gamma}'(\boldsymbol{GRADE}_{ijk} \times TIME) + \boldsymbol{\beta}'\boldsymbol{RACE}_{ijk} + \nu_k \quad (3)$$

where parameter vector $\boldsymbol{\beta}'$ represents the main effect of race on the log-hazard-rate of first suspension, thereby permitting the *level* of first-suspension risk to differ by *RACE,* with White students as the reference category. I also tested whether the profile of risk for students of different race/ethnicity differed in shape, by including interactions between race/ethnicity and time, and I retained these terms if they were required. Additionally, I included interactions between grade-level and race/ethnicity, and I retained these terms if they were required.

Findings

On virtually every school day of the academic year, in the geographic region that I studied, a relatively small number of middle-school students were suspended from school for the first time of the school year. In fact, in these sample schools, only five school days were entirely free of first suspensions over the course of the academic year: the first two days of school, the last day of school, and two other days throughout the 2009-10 school year. On average, over the entire year, the number of first suspensions

was approximately 5 per school day. But, the number of *first* suspensions per day in the region ran from a minimum of zero, on five school days, to a high of 18 suspensions, on one particularly "hazardous" day for students to experience their first suspension, in mid-September.

In Table 1, I present the results of fitting Cox-regression models describing how the occurrence of first suspension depends on the passage of time (in school days), student grade level, and student race/ethnicity. For each model, I include parameter estimates, associated standard errors, the corresponding estimated hazard ratios, and approximate *p*-values. To account for the clustering of students within schools, I provide an estimate of the between-school correlation, θ, in row 11. So, all of the hazard ratios for the parameters in each model are conditional on the estimated value of ν_k, which is identical for each student in school *k*. In row 13, I provide the log-likelihood goodness-of-fit statistic for each model. And, in rows 14 and 15, I provide the log-likelihood fit statistic for each model, and the results of a general linear hypothesis (likelihood-ratio) test in which I compare the fit of each model with that of the previous model.

Table 1. *Fitted Cox-regression models, with estimated slope parameters, standard errors, estimated hazard ratios and approximate p-values, describing how the occurrence of first suspension depends on time and race/ethnicity of the student ($n_{students}$=13,256; $n_{suspensions}$=897; $n_{schools}$=31).*

	Model A			Model B			Model C		
	Parameter	*s.e.*	Hazard Ratio	Parameter	*s.e*	Hazard Ratio	Parameter	*s.e*	Hazard Ratio
1. GRADE6 *(reference)*	-	-	1	-	-	1	-	-	1
2. GRADE7	0.1475	*0.0854*	1.1590~	0.3668	*0.1662*	1.4431*	0.5323	*0.1664*	1.7030**
3 .GRADE8	0.2054	*0.0850*	1.2281*	0.5952	*0.1625*	1.8134***	0.7433	*0.1626*	2.1028***
4. GRADE7xTIME				-0.0025	*0.0016*	0.9975	-0.0022	*0.0016*	0.9978
5. GRADE8xTIME				-0.0046	*0.0016*	0.9954**	-0.0039	*0.0016*	0.9961*
6. ASIAN							-0.8086	*0.2758*	0.4455**
7. BLACK							2.3106	*0.0821*	10.0809***
8. HISPANIC							0.4579	*0.1447*	1.5808**
9. NATIVE AMERICAN							0.5022	*0.5040*	1.6524
10. WHITE *(reference)*							-	-	1
11. θ	1.1283	*0.2912*		1.1280	*0.2912*		0.6360	*0.1870*	
12. LR test of $\theta = 0$	646.00	*p<.001*		645.99	*p<.001*		252.71	*p<.001*	
13. Log-likelihood	-8152.322			-8148.226			-7669.877		
14. LR Test of Model B vs. Model A				8.192	*p=0.02*				
15. LR Test of Model C vs. Model B							956.698	*p<.001*	
16. χ^2 test of proportional hazards	8.41 *(2)*	*p=0.02*		4.02 *(4)*	*p=0.40*		11.42 *(8)*	*p=0.18*	

17. ~ *p<0.10;* * *p<0.05* ** *p<0.01* *** *p<0.001*

RQ1: How does the risk of first suspension differ for middle-school students, over the course of a school year?

I first fitted equation 1 as Model A to estimate differences in hazard rate among grade levels, while retaining an assumption of proportionality among the fitted hazard-rate profiles. In this model, I estimate that the hazard rate for 8th grade students was 23% higher than for 6th grade students, and that the difference was statistically significant. The hazard rate for 7th grade students was 16% higher than for 6th grade students, and of marginal statistical significance. (Notice that when interpreting these anti-logged hazard ratios, I assigned a value of "1" to the reference category—6th grade.)

Furthermore, all models included the random effect of school, υ_k. Thus, in Model A, I estimated θ, the inter-school variation in hazard rates (as assumed to be gamma-distributed) was 1.13. I note then that the overall suspension rate in this sample varied substantially by school. Three middle schools recorded no suspensions for the entire year. While at the other extreme, three schools recorded overall suspension rates (*whether* students were suspended, or not) of more than 17% of the student population.

After fitting Model A, I noted a significant violation of the modeling assumption that hazard-rate profiles must be *proportional* among grades. In row 16, I supply the results of a diagnostic test of the assumption that the listed hazard rates for grades 6th through 8th grade (represented by included predictors GRADE7 and GRADE8) are proportional to each other, by inspecting the slope of the values of the Schoenfeld (1982) residuals versus time (χ^2=8.41, *df*=2, *p*=0.02), and detected a violation of the proportional-hazards assumption built into Model A. So I addressed this issue by including the two-way interaction of *GRADE* with *TIME*, specified in equation 2, and I

list the fitted model as Model B. Including the two *GRADE* × *TIME* interaction terms alleviated violations of the proportional-hazard assumption (χ^2=4.02, *df*=4, *p*=0.40) and provided an interesting insight into the effect of *TIME* on the differences in hazard rate among students in different grades.

In Model B, the population profile of risk of first suspension differed substantially among the prototypical 6th, 7th and 8th grade students. On average, the *initial* hazard rate of first suspension of 7th grade students was 44% higher than the comparable risk of suspension for 6th grade students. And, the initial hazard rate first suspension of 8th grade students was 81% higher than comparable risk for 6th grade students. Both of these estimated hazard rates, by *GRADE*, are estimates of the risk of first suspension *at the beginning of the school year.* However, I also note that the two hazard ratios for 7th and 8th grade students, compared to 6th grade students, decreased throughout the school year.[11] For example, I estimate that, on each school day, the fitted hazard rate for 8th grade students was approximately 99.5% of the hazard rate for these students on the previous day. Consequently, while the hazard rate for 8th grade was 81% higher at the beginning of the school year, compared to 6th grade students, during the school year this difference diminished as a function of *TIME*. But by the end of the school year, the risk of first suspension was 20% *lower* for 8th grade students, compared to 6th grade students.

In Figure 1, I present a plot of the estimated hazard rates versus time (in school days), by grade, incorporating both the slope estimates associated with grade and the

[11] Although the hazard rate for the interaction between 7th grade students and *TIME* was not statistically significant, the rate fell consistently between 6th and 8th grade in both level and in the direction of the interaction with time. Consequently, I retained this interaction term for the entire analysis.

baseline hazard rate profile obtained using the Kaplan-Meier (1958) strategy, and employing a kernel smoother. Thus, this plot incorporates, visually, the effect of both the level of risk (the hazard ratio) and the interaction of risk with *TIME* for 7th and 8th grade students. Each of the three fitted profiles plotted in the figure reiterate the basic shape of the underlying baseline risk, but they differ in their orientation, by grade. Notice that the obtained empirical hazard rate—and the corresponding risk of first suspension—is *not* constant over the school year. Indeed, the variability in hazard rate for all three grades appears somewhat cyclic over time.

The risk of first suspension was highest near the beginning of the school year for 8th grade students, and also, to a lesser extent, for 7th grade students. In particular, during the month of October—school day number 22 through school day number 41—a particularly high rate of first suspension was evident. On average, 7.6 middle-school students in the sample experienced their first suspension of the school year each school day during October, including 38 6th grade students, 45 7th grade students, and 69 8th grade students. This average hazard rate for October was substantially higher than during any other month of the school year.

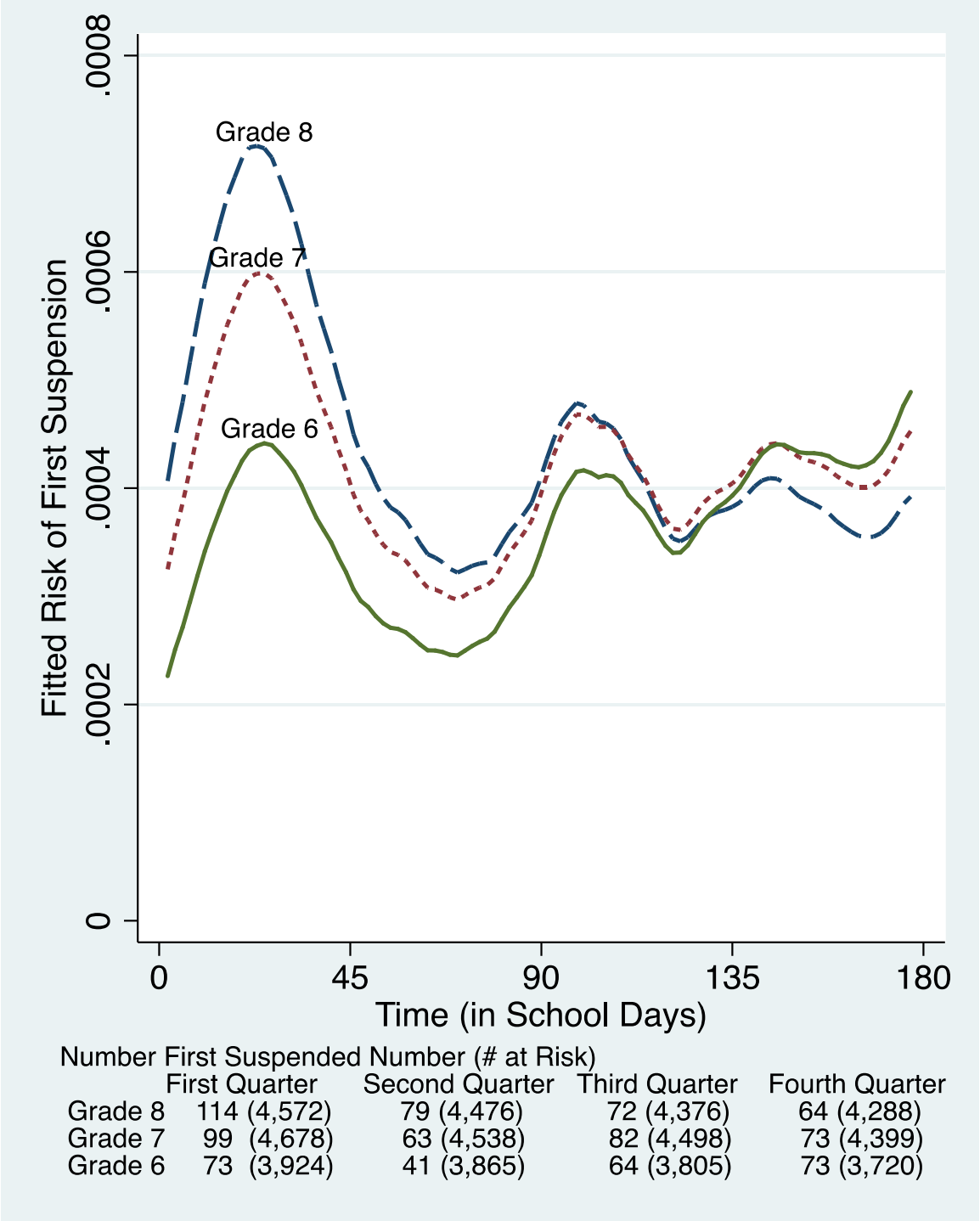

Figure 1. Estimated risk of first suspension versus time (school days), by student grade, from fitted Model B *($n_{students}$=13,256, $n_{schools}$=31, $n_{first\ suspensions}$=897).*

Interestingly, the period of lowest risk of first suspension occurred immediately after the end of the first academic quarter. The substantial "trough" in the hazard rate coincided, roughly, with the 45 school days from early November through the last week of January. During those three months, about 4.3 first suspensions occurred during each of those days—substantially lower than the total average daily rate of 5 suspensions per day. The second peak, which occurred around school day 100 (in early February), and the second trough, which occurred around school day 120 (in early March), described hazard rates that were closer to the average daily risk. An average of 6.2 first suspensions occurred in February, and 4.2 first suspensions occurred in March.

At the bottom of Figure 1, I include a risk table, displaying the number of students who were first suspended (and the number of students at risk), by academic quarter. I have disaggregated the risk table by grade level. Note that the sample of 8th Grade students had the most pronounced risk of first suspension at the beginning of the year, and then substantially diminished. Meanwhile, in contrast, the risk of first suspension for 6th grade students was highest during the 1st and 4th academic quarters, while at students were at lower risk during the 2nd and 3rd quarters. Also, notice that the quarterly summaries document that the number of first suspensions of 8th grade students during the 4th quarter was indeed lower than the number of first suspensions for 6th grade students.

Overall, I concluded that the timing of first suspensions for much of the latter part of the school year was essentially *random,* regardless of grade, as Lortie (2009) described. In fact, from the middle of the school year until its end, the hazard estimate was relatively constant, within each grade. A larger number of students in the sample who were suspended from school at all, were suspended during the first quarter—and the

risk of first suspension was substantially lower during the second quarter than during the first quarter. Still, the pattern of risk for the third and fourth quarters was steadier.[12] Also, note that the estimates of risk by grade level cross late in the school year. At the end of the school year, the risk of first suspension for 8th-grade students was *lower* than the risk for 6th-grade students.

RQ2: Is the annual temporal profile of risk of first suspension higher, and of a different shape, for Black and Hispanic students than for White students?

In my sample, overall differences by race/ethnicity, in *whether* students were first suspended or not are stunning: 34% of Black students in the sample were suspended, compared to 2% of Asian students, 6% Hispanic, 6% Native American, and 3% White, during the period of time in which each student was at risk. In Model C, I added the vector *RACE* to the previous model, as specified in equation 3. My results, again reported in Table 1, confirm the enormous disparity in the risk of suspension by student race/ethnicity. I found that Black students were at more than 10 times the risk of first suspension, and Hispanic students were at nearly 60% greater risk, compared to the reference category—White students. Asian students, as a demographic group, were at substantially lower risk than any other group. Compared to White students, I found that the risk of first suspension for Asian students was less than half the risk for White students. My estimate of the risk for Native American students was not significantly different than for White students, likely because the sample of Native students attending

[12] A keen observer might notice that there was a discrepancy between the number of students at risk at the beginning of each quarter, less the number suspended, compared to risk set for the next quarter. Recall that a few students in each quarter were *censored* at various points throughout the year, likely because they moved to another school.

public schools was quite small, and consequently I had little statistical power to detect the difference.

In Model C, I conducted a global test of the assumption of proportional hazards. I found no statistically significant evidence that this assumption was violated in Model C (χ^2=11.42, *df*=8, *p*=0.179). I found that the parameter estimates for *GRADE* were somewhat higher, and that parameter estimates associated with the effect of the *GRADE* × *TIME* interaction remained statistically significant, following the introduction of the main effects of *RACE.* I also tested whether the main effect of *RACE* interacted with *GRADE.* I found no statistically significant evidence to support this more complex model (χ^2=9.23, *df*=8, *p*=0.323). Controlling for *RACE,* I concluded that 8th grade students were at more than twice the risk of first suspension than were 6th grade students at the beginning of the school year, and that the risk for 7th grade students was about 70% higher than 6th grade students.

Finally, regarding estimates of school-level differences in students' risk of first suspension, I note, in Model C, that my estimate of inter-school variance in hazard rate, θ, was 0.64. This estimate was about half the estimate of θ for Models A and B. So, differences in student *RACE* accounted for a substantial portion, but not all, of the inter-school variation, suggesting that school segregation by race/ethnicity is an important explanation of the risk of first suspension.

In Figure 2, based on the parameter estimates from fitting Model C, I present fitted survival plots for four prototypical students, identified by their race/ethnicity. (Due to the small number of Native American students in my sample, I did not present the corresponding fitted survival profile for these students in the figure.) Each of the

prototypical students is of "average" grade level—a combination of 6th, 7th, and 8th grade. And each prototypical student attended an "average" school, with ν=1. At the beginning of the school year, all of the students are at risk of experiencing their first suspension, yet (obviously) none of the students have experienced their first suspension of the school year prior to the beginning of the school year.

Notice, first, in Figure 2, that most Asian, White, and Hispanic students "survived" the year without being suspended from school. The low level of incidence of first suspension for these students is quite striking. Consequently, it is difficult to discern the "seasonal" patterns of risk described in Figure 1. For example, for White students, at the end of the first academic quarter (school day #45), the fitted survivor function for a prototypical White student is still more than 99.1%—even after surviving the relatively "hazardous" month of October.

However, what I highlight most vividly in Figure 2 is that my estimate of the risk of first suspension for Black students is more than 10 times that for White students. Consequently, the fitted survivor plot for Black students provides a stark contrast with the fitted survivor plots for Asian, Hispanic, and White students. At the end of the first academic quarter, only 92% of the Black students in the sample have "survived" the quarter. 8% of Black students have already experienced their first suspension of the school year by early November. Even the statistically significant difference in hazard ratios between Hispanic students and White students are dwarfed in Figure 2 by the enormous disparities in rates of first suspension for Black students. And this disparity, by race/ethnicity—particularly for Black students—continued throughout the school year.

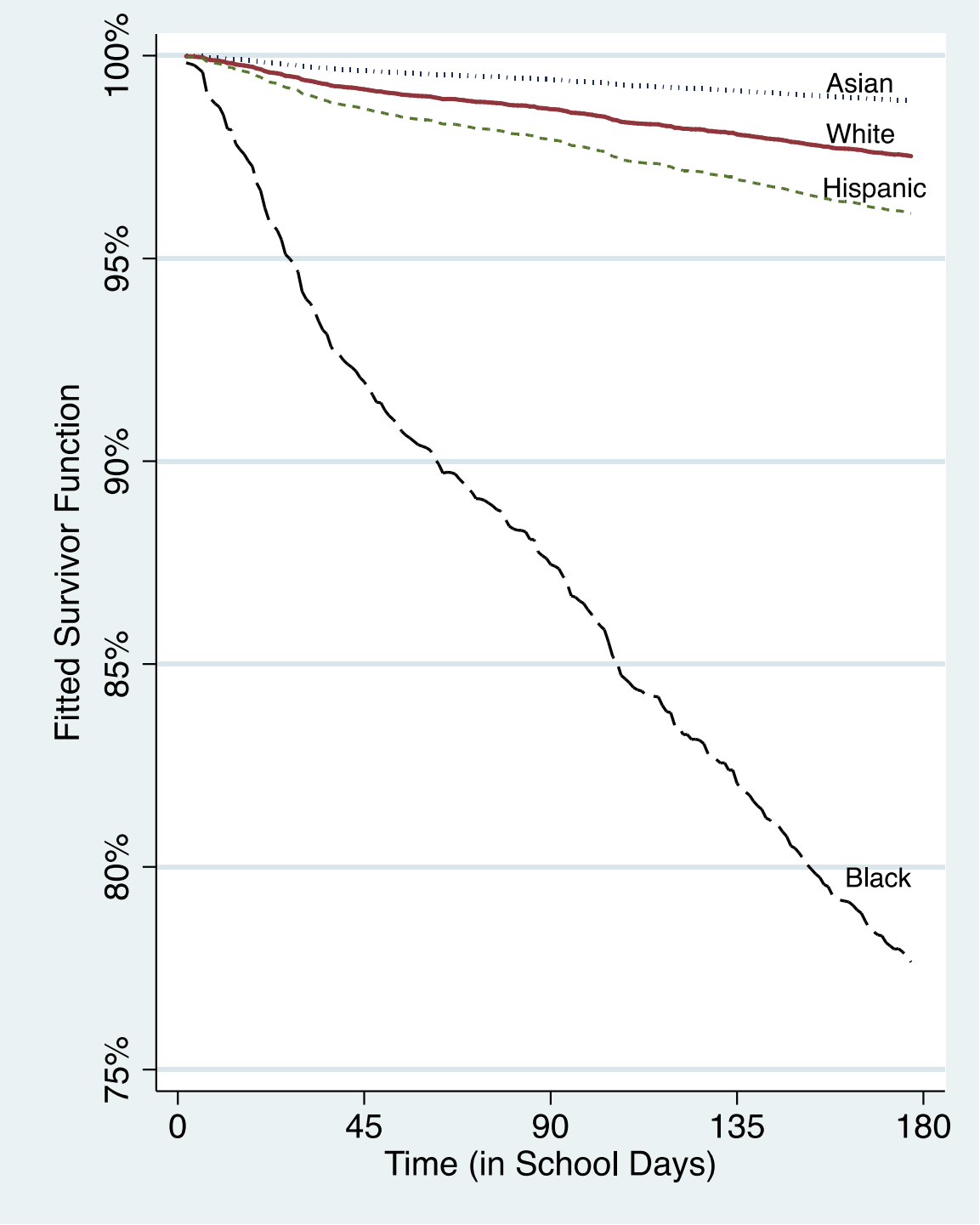

Figure 2. Estimated risk of first suspension versus time (school days) for four prototypical middle-school students, at an average ($\theta = 1$) sample school, by student race/ethnicity, from fitted Model C *($n_{students}$=13,256, $n_{schools}$=31, $n_{first\ suspensions}$=897).*

Threats to Validity

My study highlights two main stories: that the risk of first suspensions is higher at the beginning of the year, but only for students who are in 8th grade, and to a lesser extent to 7th grade; and that the risk for Black students is substantially higher than for White students. However, in this study, I describe patterns of first suspension in only one region of the country. Therefore, I cannot generalize my findings to public schools in other locations. There are likely aspects of the demographic and regional characteristics of site of this case study that influence the findings directly. Thus, I am cautious about making claims regarding the external validity of this research. In addition, in this study, I describe the risk of suspension for middle-school students during the first years of the Obama administration, before contentious debates about collective bargaining for public-school teachers began in earnest. Patterns in school suspension may be different now.

In this research, I was forced by limitations in my data to focus on the risk of first suspension over one complete year of middle school, within each of three grades. My dataset did not include any information about students' previous suspension record. For instance, I did not know, and could not incorporate into my analyses, whether a student documented as being "first suspended" in 8th grade had actually been suspended earlier, in the 6th and 7th grades. Consequently, my results may be substantially different than results about the risk of first suspension, if I had been able to incorporate the discipline history for students attending the same school during previous years into account. To reconstruct a profile that describes the risk of first suspension across an *entire* middle-school career would require three years of longitudinal data, as I would need to follow each student from their entry in 6th grade through the end of middle school. Then, in

addition, I could account for multiple occasions of suspension. I plan to adopt this approach in my future thesis research, with new longitudinal data.

Another threat to the validity of the current analysis was that the students in this sample attended middle schools with substantially different suspension rates, by school. As I noted earlier, these frailty effects make a statistically significant contribution in all of my fitted models. But does this imply that a few schools with very high suspension rates drove my results? In order to test the sensitivity of my findings to this clustering, I compared the results of fitting Model C with those from fitting the same model, but having dropped the three schools with the highest estimated value of the parameter modeling the school-level frailty, υ. The results of this sensitivity analysis were very similar with those obtained in the full sample. I also refitted Model C after dropping the three schools with the lowest frailty (the three schools recording no suspensions for the year). But again, my parameter estimates were remarkably robust, and well within the confidence intervals of the full model.[13]

Interestingly, perhaps a more basic question about my case study asks *why* students transition from elementary school to middle school at all?[14] Schwerdt and West (2013) posed this question in a study in which they estimated the impact student

---

[13] One of the schools in the sample likely failed to document how student discipline was enforced. While no students were suspended during 2009-10, during the 2008-09 school year 29 students (6%) were suspended. And, during the 2010-11 school year 35 students (7%) were suspended.

[14] For many years, the public schools in Cambridge, Massachusetts were configured as K-8 schools feeding into one large high school: Cambridge Rindge and Latin. The new configuration of twelve K-5 schools feeding into five 6-8 middle schools was implemented in the fall of 2012 http://www3.cpsd.us/ia2/ia. Conversely, in Vermont (where I presently reside), the majority of public elementary schools are configured as K-6 schools. And, many of these elementary schools feed into a single 7th -12th grade school in each "Supervisory Union."

achievement, student absences, and grade 10 dropout rate of attending public schools with different grade configurations. Similar to the case in my sample, Schwerdt and West (2013) noted that about 88% of the public-school students in Florida attended 6th through 8th grade middle schools. They concluded that entering middle school was associated with a drop in student achievement, an increase in student absences, and a subsequent higher rate of dropout rate by grade 10 (Schwerdt & West, 2013).

In light of their findings, I conducted a sensitivity analysis to investigate first suspension in the five middle schools in the sample (in five different small school districts) that were configured as 7th & 8th grade middle schools, rather than the more common 6th-8th grade middle schools. Four of the five 7th & 8th grade middle schools had very low suspension numbers. Thus, I am reluctant to draw firm conclusions, because for these four schools, each recorded less than 5 suspensions per grade level for the year. The other 7th & 8th grade school, apparently, had a substantial "incident". Nine 7th grade students were all suspended for their first time of the school year on the same day in April 2010. This incident alone accounted for more than a third of the total number of first suspensions for the school year at that grade. Perhaps these 7th graders in 7-8 schools behaved more like 6th grade students in the other 6-8 middle schools—being relatively complacent for most of the school year only to wreak havoc once spring came along? This anomaly about behavior in 7-8 schools also warrants further investigation using multi-year data. And perhaps a larger sample, including several metropolitan regions with more junior high schools, might shed some light on this issue.

Finally, an important threat to the validity of my findings concerns how the structure of my dataset itself. Note that in Model B, I included an interaction between

GRADE and *TIME,* and produced Figure 1—which did not account for the effect of *RACE.* However, the "seasonal" variation in risk revealed in this figure was driven primarily by the high rates of first suspension among Black students, compared to other students in the sample. If I had chosen to answer the second research question first, and then refined my results by subsequently including the effect of *GRADE,* I would have chosen to highlight the effect of *RACE* first—and noted that the risk of first suspension for Black students, compared to White students, was highest at the beginning of the school year. Would I have concluded that the differential effect of *TIME* on different groups pointed to an interaction between *RACE* and *TIME*, rather than between *GRADE* and *TIME?*

In order to test this rival hypothesis, I reversed the order of predictor inclusion and refitted the critical models. Thus, I first fitted a model, in which the profile of risk was hypothesized to depend only on the main effects of *RACE:*

$$\ln h(t_{ijk}) = \ln h_0(t_{ij}) + \boldsymbol{\beta' RACE}_{ijk} + \nu_k \quad (4)$$

I tested whether the proportional-hazards assumption was violated, and I concluded that it was not (χ^2=5.92, *df*=4, *p*=0.20), and that the parameter estimates for the four *RACE* categories were nearly identical.

Then, I added the effect of *GRADE* to this model, in which the profile of risk was hypothesized to depend on only the main effects of *RACE* and *GRADE:*

$$\ln h(t_{ijk}) = \ln h_0(t_{ij}) + \boldsymbol{\beta' RACE}_{ijk} + \boldsymbol{\alpha' GRADE}_{ijk} + \nu_k \quad (5)$$

In this model, I concluded that there is *marginally* statistically significant evidence of a violation of the assumption of proportional hazards in this main effects model (χ^2=12.28, *df*=6, *p*=0.056). My detailed analysis of this test revealed *two* explanations. First, I noted

that the parameter α_2, which described the impact of a student being in 8th grade, compared to 6th grade, was responsible for a large portion of the global test statistic (χ^2=6.86, *df*=1, *p*=0.01). However, the parameter β_2, describing the impact of a student being Black rather than White *also* warrants further investigation (χ^2=5.48, *df*=1, *p*=0.02), in a subsequent study, with a larger sample.

I also tested whether including an interaction between *RACE* and *GRADE*, added to the above model, specified in equation 5, might prove informative. But when I conducted a likelihood-ratio test comparing this model with the previous, the additional *eight* interaction terms (four comparison *RACE* parameters, multiplied by two comparison *GRADE* parameters) did not improve the model statistically (χ^2=10.11, *df*=8, *p*=0.26).

As I explained in my comments on **Differences in Suspension From School By Race or Ethnicity**, the demographic composition of public schools is rapidly changing, with most schools enrolling increasing numbers of Asian and Hispanic students. Consequently, I argue that simple comparisons in risk between Black students and White students are likely inadequate for future studies. And comparing the risk of suspension for Black students to non-Black students or "minority" students with Whites will probably obscure more differences than such a study might reveal. Thus, I am indeed wary about making broad claims about my research. Modeling the risk with even two categorical variables forces me, and other researchers, to make choices about the primacy of the variables—particularly when ethnicity and race are the subject.

Discussion

As I noted in my introductory comments, the work of school principals is unpredictable in many respects. Urgent demands—including vigilant attention to order and discipline—require school principals to weigh both the facts of the situation *and* the policies and procedures that serve as guides to suspending students. In this context, using survival analysis to simply describe the rates of out-of-school suspensions may seem rather banal, compared to other pressing matters. Still, I believe it somewhat puzzling that no research in the scholarly literature has tested whether or not the risk of school suspension was highest at the beginning of the school year.

My results suggest that the risk of first suspension of the school year was indeed higher at the beginning of the school year, *for 8th and 7th grade students*. For example, in Figure 1, I demonstrate that, once the school year was fully underway (after several weeks), principals were now seemingly fully engaged in maintaining order and discipline. As the hazard rate peaked during the month of October, the use of the "valuable educational device" that Justice White championed seemed to play a substantial role in preserving order and discipline in schools. Once the behavioral expectations were set during the opening days of the school year, it apparently was time for principals to "bring the hammer down"—dealing with disciplinary issues in a more decisive and formal way.

Still, *randomness* in the risk of first suspensions played a larger role than I expected. The risk of first suspensions, after an initial flurry, seemed to settle into a constant hazard rate. Thus, this case study provided evidence that the *first* out-of-school suspension for many students was imposed rather randomly throughout the school year.

No formal hypotheses about why the risk of first suspension was lower in November and December seemed apparent to me. Nor, why, perhaps, did the risk of first suspension apparently rise again in February? Consequently, using Cox (1972) regression—with a baseline hazard rate that undulated with no recognizable pattern—seemed like a justifiable modeling strategy to me.

The risk of first suspension estimated at the beginning of the school year was far more pronounced for 8th grade students than for 6th grade students. As I noted in my previous **Threats to Validity** section, this case study highlights the first suspension for students in the sample during *one school year*. It may have been that many students—particularly 7th and 8th grade students—who experienced their first suspension of the 2009-10 school year, had also been suspended the previous year. Indeed, many of these 7th and 8th grade students were likely suspended multiple times, or suspended late the previous year, perhaps continuing a pattern of misbehavior that had not been attenuated by prior administrative actions. I could not test such hypotheses in these data, but I hope to in my future thesis using longitudinal data on students over the entire three-year period of middle school.

Also, middle-school principals were working with different teams of teachers, in different grades. Theriot and Dupper (2010) highlighted the substantial difference in rates between 5th and 6th grade in their study, which focused exclusively in a school district where all of the students transitioned from elementary to middle school at that age. But in addition, their research also noted the high degree of subjectivity in "deciding" most disciplinary "cases." This subjectivity might contribute to the differences in the *profile* of risk, by grade level. For example, in this case study, many 6th grade teachers might have

chosen to teach 6th grade rather than older middle-school students. And certainly these teachers also played some part in deciding which students were sent to the principal for discipline, based on teachers' own attitudes about student behavior. Individual teachers likely contributed their own take on how 6th graders should be disciplined. Consequently, differences in staff attitudes about younger students, then, seem to reinforce Lortie's (2009) assertion that students are only gradually socialized into school behaviors. So, in my study, the risk of first suspension for 6th grade students remained relatively constant until the end of the school year.

As my study confirms, large differences in patterns of suspension are demonstrated by differences among school principals and district suspension policies, rather than by individual student behaviors—findings that have been reported for decades (Edelman et al., 1975; Skiba et al., 1997; Wu et al., 1982). Furthermore, although including the main effect of *RACE* in Model C substantially attenuated the inter-school variability, it did not eliminate it. Teacher perceptions, administrative structure, and institutional procedures and biases (Wu et al., 1982) likely still played a substantial role in *whether* students were suspended from school.

The "full" fitted model, presented as Model C and Figure 2, highlighted what I believe was the most provocative insight—that the risk of first suspension for Black students was substantially higher than any other racial/ethnic category—a full order of magnitude higher in risk, compared to White students. My research adds to a large body of studies that has researched the gap in disciplinary outcomes between Black and White students. This racial disparity has been on the agenda of researchers since Edelman (1975) and her colleagues brought this issue to the fore 40 years ago.

Future Research

Skiba et al (2011) noted that "despite widespread beliefs to the contrary, there is no previous evidence that the overrepresentation of African American or Latino students in school disciplinary outcomes can be fully explained by individual or community economic disadvantage" (p. 103). In this analysis, I present only a description of the *problem*: that Black students—and Hispanic students to a much lesser extent—are overrepresented in school suspensions. With a hazard rate for Black middle-school students of more than *10 times* the rate for White students, I believe it seems unlikely to be full explained by poverty. Thus, in my future research, I propose to parse the mediation of *RACE* by measures of student poverty like free or reduced-price lunch status.

Differences in rates of school suspension by student gender are also a potential focus of my future research. Consistently, since the 1970s, studies that have examined gender differences in school punishment have reported a substantially higher rate of school suspension for male students, compared to female students. A survey conducted by the Children's Defense Fund reported that national suspension rates were 5.4% for boys and 3.4% for girls (Edelman et al., 1975). Raffaelle and Knoff (2003) investigated the rate of out-of-school suspensions for students attending a large Florida district in 1996-97, reporting that 32% of male middle-school students were suspended at least once that year, compared to 16% of females. And, in a recent analysis of discipline data for the 2009-10 school year from a school district in Wisconsin, Sullivan, Klingbeil, and Van Norman (2013) reported that the odds of out-of-school suspension for male students were

more than twice the odds of out-of-school suspension for female students. Consequently, student gender will be an important topic when I refine and expand this analysis.

Furthermore, few other studies have explored the interaction between gender and race/ethnicity on school discipline. Ferguson (2000) investigated how Black *male* students are particularly vulnerable to strict administration of discipline, like school suspensions, compared to White male students. And Morris (2005) observed, in a case study of a Texas school, that Black *female* students were likely to be disciplined at a rate similar to Black *male* students. Thus, in my further research, I plan to explore how the effect of *RACE* may itself be moderated by gender.

Conclusion

Using continuous-time survival analysis, I concluded that the risk of first suspension for middle-school students was higher at the beginning of the school year for 8th and 7th grade students. However, other than this, the risk of first suspension was relatively flat, holding steady at about five incidents of first suspensions per day, among the students in the 31 schools during the second half of the school year. Employing Cox-regression analysis, I also found that the risk differed by grade in school. At the beginning of the school year, the risk of first suspension for 8th grade students was more than double the risk for 6th grade students. But this difference in risk between grades diminished over time. Additionally, the risk of first suspension for Black students was substantially higher—more than 10 times the risk of first suspension for White students.

References

Anderson, J. D., (1988). *The education of Blacks in the south, 1860-1935*. Chapel Hill: University of North Carolina Press.

Arcia, E. (2007). A comparison of elementary/K-8 and middle schools' suspension rates. *Urban Education, 42*(5), 456-469. doi:10.1177/0042085907304879

Arum, R. (2003). *Judging school discipline: The crisis of moral authority.* Cambridge, Mass.: Harvard University Press.

Arum, R., & Preiss, D. (2009). Law and disorder in the classroom. *Education Next, 9*(4)

Beck, A. N., & Muschkin, C. G. (2012). The enduring impact of race: Understanding disparities in student disciplinary infractions and achievement. *Sociological Perspectives, 55*(4), 637-662. doi:10.1525/sop.2012.55.4.637

Butts, R. (1955). Our tradition of states' rights and education. *History of Education Journal, 6*(3), 211-228.

Cleves, M. A., Gould, W. W., Gutierrez, R. G. (2010). *An introduction to survival analysis using Stata (3rd. ed.)*. College Station, Tex.: Stata Press.

Cox, D. R. (1972). Regression models and life-tables (with discussion). *Journal of the American Statistical Association, Series B, 34*, 187-220.

Dugan, L., Lafree, G., & Piquero, A. R. (2005). Testing a rational choice model of airline hijackings. *Criminology, 43*(4), 1031-1066.

Edelman, M. W., Beck, R., & Smith, P. V. (1975). *School suspensions--are they helping children?* Cambridge, Mass.: Children's Defense Fund.

Efron, B. (1977). The efficiency of Cox's likelihood function for censored data. *Journal of the American Statistical Association, 72*(359), 557-565.

Ferguson, A. A. (2000). *Bad boys: Public schools in the making of Black masculinity.* Ann Arbor, MI: The University of Michigan Press.

Goldin, C. D., & Katz, L. F. (2008). *The race between education and technology.* Cambridge, Mass.: Belknap Press of Harvard University Press.

Goss v. Lopez, 419 U.S. 565, *Justia US Supreme Court Center* (U.S. Supreme Court 1975).

Gregory, A., Cornell, D., & Fan, X. (2011). The relationship of school structure and support to suspension rates for black and white high school students. *American Educational Research Journal, 48*(4), 904-934.

Heikinheimo, T., Broman, J., Haapaniemi, E., Kaste, M., Tatlisumak, T., & Putaala, J. (2013). Preceding and poststroke infections in young adults with first-ever ischemic stroke effect on short-term and long-term outcomes. *Stroke, 44*(12), 3331-3337. doi:10.1161/STROKEAHA.113.002108

Hinojosa, M. S. (2008). Black-white differences in school suspension: Effect of student beliefs about teachers. *Sociological Spectrum, 28*(2), 175-193. doi:10.1080/02732170701796429

Kaplan, E. L., & Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association, 53*(282), 457-481.

KewalRamani, A., Gilbertson, L., Fox, M. A., Provasnik, S. (2007). *Status and trends in the education of racial and ethnic minorities.* NCES 2007-039. National Center for Education Statistics.

Kinsler, J. (2011). Understanding the black-white school discipline gap. *Economics of Education Review, 30*(6), 1370-1383. doi:10.1016/j.econedurev.2011.07.004

Lomawaima, K. T. (1999). *The unnatural history of American Indian education.*

Lortie, D. C. (2009). *School principal: Managing in public.* Chicago: University of Chicago Press.

Losen, D., & Skiba, R. J. (2010). *Suspended education: Urban middle schools in crisis.* Southern Poverty Law Center.

Martin, J. (2014, April 20 2014). As G.O.P wedge, the common core cuts both ways. New York Times, pp. A1.

McCarthy, J. D., & Hoge, D. R. (1987). The social construction of school punishment: Racial disadvantage out of universalistic process. *Social Forces, 65*(4), 1101-1101-1120.

Morris, E. W. (2005). "Tuck in that shirt!" Race, class, gender, and discipline in an urban school. *Sociological Perspectives, 48*(1), 25-48.

National Governors Association Center for Best Practices & Council of Chief State School Officers. (2010). Common core state standards. Washington, DC.

Nichols, J. D. (2004). An exploration of discipline and suspension data. *The Journal of Negro Education, 73*(4), 408-408-423.

No child left behind (NCLB) act of 2001, Pub. L. No. 107-110, § 115, Stat.1425, (2002).

Office of Management and Budget. (1997). Revisions to the standards for the classification of federal data on race and ethnicity. Retrieved February 21, 2014, from www.whitehouse.gov/omb/fedreg/1997standards.html

Omi, M., & Winant, H. (1994). *Racial formation in the United States: From the 1960s to the 1990s (2nd ed.)*. New York: Routledge.

Ornstein, A. (1982). Student disruptions and student rights: An overview. *The Urban Review, 14*(2), 83-91.

Payne, A. A., & Welch, K. (2010). Modeling the effects of racial threat on punitive and restorative school discipline practices. *Criminology, 48*(4), 1019-1062. doi:http://dx.doi.org.ezp-prod1.hul.harvard.edu/10.1111/j.1745-9125.2010.00211.x

Petras, H., Masyn, K. E., Buckley, J. A., Ialongo, N. S., & Kellam, S. (2011). Who is most at risk for school removal? A multilevel discrete-time survival analysis of individual- and context-level influences. *Journal of Educational Psychology, 103*(1), 223-237. doi:10.1037/a0021545

Raffaele Mendez, L. M., & Knoff, H. M. (2003). Who gets suspended from school and why: A demographic analysis of schools and disciplinary infractions in a large school district. *Education and Treatment of Children, 26*(1), 30-30-51.

Race to the top act of 2011, H.R. 1532--112th Congress (2011).

Schoenfeld, D. (1982). Partial residuals for the proportional hazards regression model. *Biometrika, 69*(1), 239-241.

Schwerdt, G., & West, M. R. (2013). The impact of alternative grade configurations on student outcomes through middle and high school. *Journal of Public Economics, 97*(0), 308-326. doi:http://dx.doi.org.ezp-prod1.hul.harvard.edu/10.1016/j.jpubeco.2012.10.002

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence.* Oxford; New York: Oxford University Press.

Skiba, R. J., Horner, R. H., Choong-Geun Chung, Rausch, M. K., May, S. L., & Tobin, T. (2011). Race is not neutral: A national investigation of African American and Latino disproportionality in school discipline. *School Psychology Review, 40*(1), 85-107.

Skiba, R. J., Peterson, R. L., & Williams, T. (1997). Office referrals and suspension: Disciplinary intervention. *Education & Treatment of Children (ETC), 20*(3), 295.

Sullivan, A. L., Klingbeil, D. A., & Van Norman, E. R. (2013). Beyond behavior: Multilevel analysis of the influence of sociodemographics and school characteristics on students' risk of suspension. *School Psychology Review, 42*(1), 99-114.

Tamura, E. H. (2001). Historiographical essay. *History of Education Quarterly, 41*(1), 58.

Theriot, M. T., & Dupper, D. R. (2010). Student discipline problems and the transition from elementary to middle school. *Education and Urban Society, 42*(2), 205-222. doi:10.1177/0013124509349583

Townsend, B. L. (2000). The disproportionate discipline of African American learners: Reducing school suspensions and expulsions. *Exceptional Children, 66*(3), 381-381-391.

U.S. Census. (2002). Measuring America: The decennial censuses from 1790 to 2000. No. POL/02-MA(RV). U.S. Department of Commerce.

U.S. Department of Education. (2008). New race and ethnicity guidance for the collection of federal education data. Retrieved January 18, 2014 from http://www2.ed.gov/policy/rschstat/guid/raceethnicity/index.html

U.S. Department of Education. (2012). *Digest of educational statistics.* Washington DC: U.S. Dept. of Health, Education, and Welfare, Office of Education.

United States Department of Education Office for Civil Rights. (2013). Civil rights data collection www2.ed.gov/about/offices/list/ocr/docs/crdc-2011-12-p1-p2.doc

Vaupel, J. W., Manton, K. G., & Stallard, E. (1979). The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography, 16*(3), 439-454.

Wallace, J.,John M., Goodkind, S., Wallace, C. M., & Bachman, J. G. (2008). Racial, ethnic, and gender differences in school discipline among U.S. high school students: 1991-2005. *Negro Educational Review, 59*(1), 47-62.

Weibull, W. (1951). A statistical distribution function of wide applicability. *Journal of Applied Mechanics,* (September), 293-297.

Wu, S., Pink, W., Crain, R., & Moles, O. (1982). Student suspension: A critical reappraisal. *The Urban Review, 14*(4), 245-303. doi:10.1007/BF02171974

Zheng, H., & Thomas, P. A. (2013). Marital status, self-rated health, and mortality: Overestimation of health or diminishing protection of marriage? *Journal of Health and Social Behavior, 54*(1), 128-143. doi:10.1177/0022146512470564

You again? Improving the Estimation of the Occurrence and Timing of Repeated School Suspensions Using Survival Analysis

Stephen L. Hoffman

Harvard Graduate School of Education

You again? Improving the Estimation of the Occurrence and Timing of Repeated School Suspensions Using Survival Analysis

During my professional career in the public schools, I served as a middle-school principal and a high-school assistant principal. On most school days, I dealt with at least a handful of students whose misbehavior warranted action on my part. Sometimes, I suspended students out of school following strict district procedures. But more frequently, my decisions followed the ad-hoc standards advocated by my immediate superiors. At the same time, I was also deeply troubled by what I perceived as the disproportionate punishment meted out to Black and Hispanic students. It is these experiences in public schools, and my subsequent reading of the associated scholarly literature, that have motivated me to analyze the consequences for children of disciplinary policies (Beck & Muschkin, 2012; Gregory, Cornell, & Fan, 2011; Kinsler, 2011; Petras, Masyn, Buckley, Ialongo, & Kellam, 2011; Skiba et al., 2011).

In my earlier research, I have focused on the substantial racial/ethnic disparities in school suspensions. In one study, for instance, I demonstrated that a sizeable difference between Black students and White students in the proportion of days suspended *increased*, following the expansion of zero-tolerance policies (Hoffman, 2014c). However, that work was limited to the analysis of publicly available data, in which the occurrence of school suspensions had been aggregated to the school level. In a second study—for my qualifying paper (QP)—I conducted a "single-spell" survival analysis of the *first* out-of-school suspension of middle-school students over one school year, using a confidential, rich, longitudinal dataset (Hoffman, 2014b). However, my QP findings were limited because I was only able to estimate the risk of students' first suspension of the

school year for separate cohorts of sixth, seventh, and eighth graders, rather than being able to document the entire profile of risk for the entire period of middle school, from sixth through eighth grade.

Still, early adolescence remains an important life stage for investigation because, during middle school, suspension rates typically exceed those of elementary-school students, and sometimes even surpass corresponding rates for high-school students (Theriot & Dupper, 2010; Losen & Skiba, 2010; Raffaele Mendez & Knoff, 2003). And so, in this paper, I extend my work further, using repeated-spells survival analysis, described by Willett and Singer (1995), and others (Cook & Lawless, 2007; Masyn, 2009; Rabe-Hesketh & Skrondal, 2012). Using this more sophisticated technique, I improve on my earlier research by describing the timing of out-of-school suspensions for individual students, followed over the entire sixth through eighth grade period, and not just for the first suspension, but for the repeated occurrence of out-of-school suspensions. Furthermore, I extend my documentation of the substantial inequities in the risk of suspension, not only by students' race/ethnicity, but also by their gender and family income.

I have organized this paper into five sections, following this introduction. First, in the *Background and Context* section, I review relevant literature on school suspension generally. I describe commonly used analytic techniques for estimating the suspension rate, which summarize *whether* students were suspended from school—especially when documenting differences in suspension rates among key demographic groups. Then, I argue that a hitherto neglected aspect of analyzing the consequences of school discipline is addressing the important question of *when* suspensions occur—and that survival

analysis is a plausible strategy for this purpose. I close this background section by stating my specific research questions. Next, in my *Research Design* section, I describe the design of this case study, including my research site, dataset, measures, and the data-analytic methods used to address my specific research questions. Third, I present my *Findings* from this analysis. Fourth, I discuss the limitations of my analysis in a section on *Threats to Validity.* Finally, I synthesize and review the meaning and significance of the findings in a fifth and final *Discussion* section.

Background and Context

For decades, school administrators have suspended students from school for disciplinary reasons (Edelman, Beck, & Smith, 1975; Wu, Pink, Crain, & Moles, 1982; Arum, 2003). During the 1960s and 1970s, the issue of school suspensions came to the forefront of public debate, partly as a result of the surfacing of racial inequities in suspension rates by the civil-rights movement (Arum, 2003). In the early 1970s, the *Office of Civil Rights* first began collecting data about suspensions in the United States (Edelman et al., 1975). This led to new policies for the administration of school suspensions, ordered by the courts and sanctioned by the federal government (Edelman et al., 1975; Ornstein, 1982).

Ultimately, the U.S. Supreme Court settled the question of *whether* students might be suspended out-of-school due to their behavior in *Goss v. Lopez* (1975). Noting that disciplinary events in schools occurred frequently, Justice White maintained that out-of-school suspension was not only "a necessary tool to maintain order, but a valuable educational device" (*Goss v. Lopez*, p. 580). This broad framework for the use of school exclusion still shapes U.S. public-school policies today (Arum & Preiss, 2009).

However, school suspensions impose substantial costs on students, their families, and the community—costs that I have observed both professionally and personally. More than 40 years ago, Edelman, Beck and Smith (1975) recommended that schools reduce the use of out-of-school suspension. They argued that patterns of suspension were largely a consequence of differences among school administrators and suspension policies, rather than different behavioral patterns among students. This was confirmed empirically in a subsequent national study of 641 public secondary schools conducted in 1976—the year after *Goss v. Lopez* (1975). Wu, et al (1982) concluded that student misbehavior was only one factor resulting in an out-of-school suspension. A principal's decision to suspend a student is also mediated by teachers' perceptions and beliefs, administrative structures, and institutional biases (Wu et al., 1982).

Recently, the federal government again underscored the overuse of suspension, particularly for Black students (U.S. Department of Education, 2014). Former Secretary of Education Arne Duncan argued that simply "relying on suspensions and expulsions, however, is not the answer to creating a safe and productive school environment" (p. i). Rather, schools must evaluate "the impact of their discipline policies and practices on all students using data and analysis" (p. iii). In this paper, I respond to Secretary Duncan's call.

Estimating the Traditional Suspension Rate

Part of understanding the impact of school suspensions is developing good measures to assess the issue. The simplest method for estimating the *suspension rate* is to count *whether* students were suspended during a defined period of time and divide this by the total number of students present at the beginning of the period. For example, a recent

review by the National Center for Education Statistics (NCES) estimated that 3,325,000 students were suspended out-of-school at least once during 2006—a suspension rate of 7% for all students in public elementary and secondary schools at the time (U.S. Department of Education, 2012).

However, when reported in isolation, this summary statistic masks important subtlety. Most research on suspensions has focused, historically, on the disparities that exist in suspension rates among students in different demographic groups, overlooking how these critical covariates might interact. Additionally, a typically unknown fraction of students have been suspended repeatedly. Thus, caveats like "at least once" obscure the full extent of the problem. I discuss each of these challenges in more detail below. I first summarize previous research about differences in the suspension rate among three critical demographic groups: student gender, race, poverty, and the interactions among these groups. I also discuss that student grade level is rarely examined in estimates of the suspension rate.

Student Gender. The old nursery rhyme muses that boys are made of "snips & snails & puppy dogs' tails," while girls are made of "sugar & spice & all things nice." Indeed, consistent with this, differences in rates of out-of-school suspensions between male and female students is one of the most salient facts about school discipline. Since the 1970s, studies that have estimated differences in the rate of suspension among student groups have consistently reported substantially higher suspension rates for boys than for girls (Edelman et al., 1975; McFadden & Marsh II, 1992; Skiba, Peterson, & Williams, 1997). Raffaelle and Knoff (2003), analyzing the suspension rate for a large Florida district in 1996-97, found that 32% of male middle-school students were suspended at

least once that year, compared to 16% of females. Similarly, analyzing discipline data from a Wisconsin school district for the 2009-10 school year, Sullivan, Klingbeil, and Van Norman (2013) also concluded that the odds of out-of-school suspension for male students were more than twice the odds of out-of-school suspension for female students.

In a recent study, using data from the Early Childhood Longitudinal Study: Kindergarten Cohort (ECLS-K), Bertrand and Pan (2013) asserted, perhaps provocatively, that "biological forces are at play" when explaining gender differences in behavior problems at school (p. 60). They concluded that differences in eighth-grade suspensions between boys and girls in the ECLS-K sample were mediated by another variable—whether students' were raised in a two-parent household or not. (Note, however, that they did not distinguish between in-school suspensions and out-of-school suspensions.) Specifically, the gender gap in suspension between boys and girls was higher for boys raised by single mothers, compared to boys raised in two-parent households. And while I do not have data about whether the students in my sample were from one-parent or two-parent households, I do have information about student race and level of poverty—two social factors that may mediate plausible gender behavioral differences. So I will test whether gender differences in risk interact with these two other critical covariates in my analysis.

Student Race. Decades of research have also documented important racial/ethnic[15] disparities in the rate at which students are suspended out-of-school (Edelman et al., 1975; McCarthy & Hoge, 1987; Skiba et al., 1997; Townsend, 2000;

[15] While race is a social construction—a mixture of ignorance and fiction—that means different things in many contexts, I believe that it is unwise to ignore our present classification system, as race matters to people who are subject to different risks.

Losen & Skiba, 2010; Kinsler, 2011; Petras et al., 2011; Skiba et al., 2011; Sullivan et al., 2013). As Skiba et al. (2011) commented, in a study of Black disproportionality in school discipline, "race is not neutral." By analyzing school disciplinary records in 364 elementary and middle schools, the authors, using logistic regression analysis, concluded that the probability that Black students would be suspended from school was greater than for White students, for the same or similar misbehavior.

In another study also focused primarily on disproportionality in school suspension by race/ethnicity, Wallace, Goodkind, Wallace, and Bachman (2008) analyzed data collected from a nationally representative survey of high-school students. Using logistic regression analysis, they estimated that the odds that Black students reported being suspended from school were more than three times the odds for White students. While these two studies differ in data-collection methodology— with Skiba et al. (2011) utilizing administrative records, and Wallace et al. (2008) using individual survey data— both studies highlight the disproportionate risk of school suspension for Black students. This disparity has changed little since the research of Edelman et al. (1975) four decades ago. And so, in this paper, I have focused in particular on differences in the risk of school suspension between Black students and White students.

Gender, Race and Poverty. Skiba et al. (2011) has argued that the over-suspension of Black students—both male and female—cannot be explained fully by students' level of poverty. Rather, differences in suspension rates by gender, race/ethnicity, and levels of poverty operate *simultaneously* to color students' school experience. A few researchers have looked at the joint effects of these demographic characteristics on the suspension rate. For example, Sullivan, et al. (2013) estimated that

the odds of out-of-school suspensions for male students were twice the odds for female students; controlling for gender, the odds for Black students were more than three times the odds for White students; and, controlling for gender and race, the odds of a school suspension for students eligible for free or reduced-price lunch (*FRPL*)—a commonly used proxy for student poverty (Aud et al., 2013)—were nearly three times the odds of those not in poverty.[16]

Note, however, that in most of these studies, researchers have focused, primarily, on suspension rates described in terms of odds-ratios, where differences are estimated as *proportional* differences. Interactive effects between race, gender, and poverty are rarely examined or discussed, thus missing an important opportunity to examine the roots of disproportionality. In particular, when researchers attempt to circumvent or control for race (e.g. Bertrand & Pan, 2013) within studies about education, they are softening an important factor for examining equity and inequity in our schools—the intersectionality of these social identities.

Recently, publicly available estimates of the suspension rate have become routine—reported by government agencies, typically, as cross-tabulated summaries. For instance, a recent *Data Snapshot* published by the U.S. Department of Education Office for Civil Rights reported that 16.4% of Black public-school students (kindergarten through 12th grade) were suspended during the 2011-12 school year, compared to 4.6% of White students, (U.S. Department of Education Office for Civil Rights, 2014). And

[16] Nationally, 47.5% of public-school students were eligible for F/R lunch during the 2009-10 school year (U.S. Department of Education, 2011), though this percentage varied considerably by state, from 23.5% in New Hampshire, to 70.7% in Mississippi and 72.3% in DC.

crossing the race and gender classifications, the government reported substantial differences in the suspension rate: 20% of Black boys and 12% of Black girls were suspended, compared to 6% of White boys and 2% of White girls. Thus, disparities in national suspension rates *differ* by gender, whether calculated as proportions (a gender differential of 1.67 for Black students, compared to 3 for White students), or as odds ratios (a gender differential of 3.9 for Black students, compared to 6.7 for White students). And the racial gaps in suspension rates, by gender, require further investigation.

Moreover, this cross-tabulation in national rates of suspension by gender and race does not control for socio-economic status, possibly clouding an interactive effect of poverty on a potential gender-race interaction. However, in a recent study, Autor, Figlio, Karbownik, Roth, and Wasserman (2015) analyzed data from Florida that includes measures of socio-economic status to explain the gender gap in educational outcomes, including suspension rates. They found that about 11% of Florida public school children were suspended for at least one day per school year during grades 3 through 8, and that suspension rates were more than twice as high for males as for females. Furthermore, the gender difference in suspension rates was twice as large for Black students, compared to White students—in conflict with the national summary cited above. But they also found that these gender differences in the suspension rate were larger for students in low-SES households—perhaps a clue to this puzzle (Autor et al., 2015). Note, too, that Autor and his colleagues estimated suspension rates only for students in grades 3 through 8. Still, it is difficult to reconcile these two different conclusions without including another dimension besides race and gender.

And so, in this paper, I first estimate the risk of out-of-school suspension for each of these three key demographic dimensions—gender, race (specifically, differences between Black students and White students), and eligibility for free/reduced-price lunch—separately, in my own investigation of the risk of suspension. Then, I explore how these three dimensions interact with each other, and with time. I examine whether the respective influences of gender, race, and poverty can be described as distinct risk factors, or whether these dimensions work together in unexpected ways. In this paper, I describe these interactions, whereby the systematic effect of each of these covariates may not be adequately predicted as simply the sum of the main effects for each of these dimensions.

Student Grade. Most summaries of the suspension rate—like those provided by the U. S. Department of Education, or the research of Sullivan, et al. (2013)—describe the average prevalence of suspension amongst all K-12 public-school students across grades. However, substantial differences in suspension rates exist among students attending different grades. But, as Losen and Skiba (2010) explained, no nationally representative reports—based on federally collected discipline data—have been disaggregated by grade-level.

However, some states *do* report suspension rates by grade. For example, in Wisconsin, rates of out-of-school suspensions have been lowest, historically, for students in kindergarten and were higher in each successive grade throughout elementary school. A substantial jump in suspensions then coincided with sixth grade—typically the start of middle school. And suspension rates remained comparatively high throughout secondary school. Suspension rates were highest in 9th grade, and were systematically lower in 10th

through 12th grades (Department of Public Instruction, 2013). In this paper, motivated by my professional experience as a secondary-school principal, as well as the findings from several studies that have noted high rates of suspension for middle-school students (Theriot & Dupper, 2010; Losen & Skiba, 2010; Raffaele Mendez & Knoff, 2003) I focus exclusively on the suspensions of students during middle school.

Refining the Investigation of Suspension and the Estimation of Suspension Rates

In recent work that is more sophisticated methodologically than simple comparisons of suspension rates, several scholars have investigated the occurrence of suspensions longitudinally, for intact cohorts of students followed across multiple school years. For instance, a report prepared by the Council of State Governments Justice Center, in partnership with the Public-Policy Research Institute at Texas A&M University, analyzed the school records of all public-school students in Texas who began 7th grade in 2000, 2001, and 2002, and were followed through the end of high school. One of their most important findings was that, between 7th grade and 12th grade, 31% of all students were suspended out-of-school at least once during secondary school (Fabelo et al., 2011). Note that this statistic—summarizing whether, *if ever*, students were suspended from secondary school—is very different from the simple suspension rate of 6.5% in Texas, estimated for the 2007-08 school year.[17]

Taking this approach a step further—although focusing only on the *beginning* of students' school careers—Petras, Masyn, Buckley, Ialongo, and Kellam (2011) investigated the occurrence of school suspensions in the public schools of Baltimore.

[17] Author's calculation based on reported rates by the Texas Education Agency. Retrieved on October 28, 2014 from ritter.tea.state.tx.us/cgi/sas/broker

Specifically, they used discrete-time survival analysis to investigate what is referred to as the *risk* of first suspension from school—that is, they estimated the conditional probability of suspension in a particular time period, given no suspensions prior to that period—across 1st through 7th grade. They focused on the risk of first suspension in a student's school career because they hypothesized that the timing of the first suspension would then be indicative of later school problems. They found that 22.8% of students were suspended at least once during those grades. They also estimated that the risk of first suspension was low in the primary grades, but substantially higher during 6th and 7th grade.

Similarly, in my qualifying paper, I employed a type of continuous-time survival analysis—Cox-regression analysis (Cox, 1972)—to describe students' risk of first suspension over the course of one school year, for separate cohorts of students in three middle-school grades. I concluded that the risk of first suspension was higher near the beginning of the school year, though students' first suspension could occur on virtually any school day (Hoffman, 2014b). Indeed, some students were even suspended for the first time of the school year in early June—a phenomenon that school principals know full well, but perhaps researchers have overlooked.

One of the limitations of my Cox-regression approach, however, was that it was only semi-parametric, with the risk of suspension being investigated only as a function of selected covariates, such as student race/ethnicity, and not directly as a function of time itself. With no explicit modeling of the baseline hazard rates, I needed to recover post-model-fitting estimates of differences in the risk of suspension throughout the year from the data themselves to address *when* the probability of an occurrence of school

suspension was at its greatest (Singer & Willett, 2003) once the impact of student race had been removed. In the current paper, I improve on this by fitting discrete-time survival models in which I incorporate explicit parameterization of the baseline risk function as a function of time. I am then able to test hypotheses about the timing of suspensions explicitly within the modeling framework.

Repeated Suspensions from School

Losen and Skiba (2010) described rule breaking as normal behavior—particularly for middle-school students, who may be accustomed to challenging authorities, at home and at school—as part of natural adolescent social development. Indeed, in *Goss v. Lopez* (1975) the Supreme Court noted that disciplinary events in schools were frequent occurrences, and counseled "immediate, effective action" including out-of-school suspension to maintain order (p. 580). But while the Court foreshadowed that students require frequent disciplinary intervention, estimates of the suspension rate, like the results of the NCES report cited earlier, estimate simply *whether* students were suspended or not during the school year, even if they were suspended once or multiple times during that year. It is true that some behavioral norms may be established early in the school year, but developmental factors continue to play a substantial role as students are socialized into complying with school norms over the year (Lortie, 2009). So, some students may be suspended a single time, or not at all, while others may be suspended multiple times.

Many published reports of out-of-school suspensions provide little or no information about whether students were suspended repeatedly during the year. However, I noted two exceptions. First, the *Data Snapshot*, cited above, estimated that 3.9% of public-school students were suspended *once* during the 2011-12 school year, while 3.2%

were suspended multiple times (U.S. Department of Education Office for Civil Rights, 2014). And in the study of Texas secondary-school students, also cited above, Fabelo, et al. (2011) found that, among students who *were* suspended, they averaged *eight* suspensions during their secondary-school career.

Making sense of published reports of historical suspension rates across multiple years produces further complications. As noted above, some states provide grade-level suspension rates each year. But when analyzing aggregate data, it is impossible to know whether students who were suspended one year were suspended again the next. For example, one school district (from the same region as the district featured in this case study) reported that 55 of the approximately 750 6th grade students were suspended, at least once, during the 2011-12 school year—a suspension rate of 7%. The following year, 69 of the 7th grade students were suspended—a suspension rate of 9%. Thus, somewhere between *none* of the 55 students suspended as 6th graders and *all* of them plus 14 others may have been suspended in 7th grade. Thus, the percentage of 7th graders suspended who had been suspended in 6th grade fell somewhere between 0% and 80%, a statistic that requires more precise estimation.

Similarly, in my previous work in which I estimated the risk of a first suspension, I concluded that this risk was higher for 8th grade students than for 6th grade students (Hoffman, 2014b). However, the 7th and 8th grade students in that study may have been suspended from middle school in previous years. Thus, estimating the proportion of those suspended among those who had been suspended in previous years is impossible without access to—and analysis of—longitudinal data spanning the middle-school grades. When such data are available, methodologists have suggested the use of a multiple-spells

approach to investigate the occurrence and timing of repeated suspensions (Willett & Singer, 1995). This innovative analytic strategy extends the standard methods of survival analysis used for the modeling of the occurrence of single events to accommodate the risks of multiple "events" occurring sequentially over time.

Employing this strategy in the context of school discipline, the "events" to be modeled are the decision of a school administrator to suspend a student out-of-school, due to her/his misbehavior. The "spells" are time-periods of students' school careers during which they were at risk of being suspended a first time, a second time, a third time, et cetera. The first spell began at the start of a defined period (for example, the first day of middle school), lasting until a student was suspended (if ever). A second spell, if it exists, began upon students' return from a first out-of-school suspension. Similarly, a third spell, if it exists, began upon students' return from a second suspension. And so, in this paper, I use repeated-spells survival analysis to describe the profile of risk of multiple suspensions experienced by students sequentially, estimating these risks over students' entire three-year middle-school career.

Conclusion: Specific Research Questions

In the above discussion, I have highlighted studies that have reported estimates of the student suspension rate for a single school year or grade. However, I have also noted that few studies have examined repeated suspensions of students—in multiple spells of suspension, sequentially over several grades—even though the risk of subsequent suspensions may be non-trivial. Nor have these studies documented whether the risk of multiple suspensions differed substantially by key student demographic dimensions, as suggested most prominently by the staggeringly high suspension rates for Black students

in American schools. To tackle this, in this paper, I conduct a quantitative case study among students in one large school district from which I have been granted access to rich individual longitudinal data on the occurrence of multiple suspensions. I address the following research questions:

1. *When* are students at the greatest risk of being suspended—either once, or repeatedly—from public middle schools? Specifically,
    a. Within a spell, is the risk of suspension higher at particular time periods, lower at others?
    b. Does the temporally-dependent profile of risk differ by spell? Specifically, is the risk *higher* in subsequent spells than the risk of a first suspension, and if so, by how much? And, net of its level, does the shape of the risk profile differ by spell?
2. Over all spells, how does the profile of risk of suspension *differ* along key demographic dimensions? Specifically,
    a. Are middle-school boys at greater risk of suspension than middle-school girls?
    b. Are Black students at greater risk of suspension than White students?
    c. Are students from low-income families at greater risk of suspension than students from more affluent families?
    d. Are the respective influences of gender, race, and poverty on the risk of suspension additive or interactive?

Research Design

Site

For this case study, I focus on students enrolled in the public middle schools of one school district to whose data I have been granted confidential access. Under the terms of my data agreement, I must conceal the name of this district to protect the privacy of the students whose middle-school careers were investigated. But, fortunately for my research focus, the district—one of the largest in its state—is an urban district, with a racially diverse and economically heterogeneous student population. Note, though, that the *statewide* rate of school suspension for White students was *lower* than the national average, but the statewide suspension rate was substantially *higher* for Black students (U.S. Department of Education, 2012).

The district is ideal for my case study for several reasons. First, it had a well-established governance structure. All of the middle-school principals reported directly to an assistant superintendent. Consequently, they were held accountable for their decisions about school suspension, both through supervision and mentorship by the assistant superintendent and through formal procedures for parents and students to protest the principals' decisions. Further, the district followed a discipline plan that included defined criteria for the out-of-school suspension of students in virtually every possible scenario, from inappropriate language through weapons violations. Second, perhaps regrettably from a substantive perspective (although not a methodological one!), the district reported a sizeable number of incidents of out-of-school suspensions during the years of my study. Third, in summary statistics, the district also reported very large differences in suspension rates by gender, by race, and by levels of poverty.

Dataset

I have the educational records of all students who enrolled in 6th grade in this district in the fall of 2007, and can follow them through the end of 8th grade in June 2010. I constructed my analytic dataset by combining information from three linked administrative datasets. The first dataset contributed student enrollment information, identifying the date that each student enrolled in a district school, and the subsequent date that s/he exited the school. The second dataset contributed student demographic information. The third dataset recorded the date of any incident requiring formal disciplinary action, such as out-of-school suspension or expulsion, and noted the number of days that a student was removed from school.

The resulting merged dataset contains a student identification number, information on the date each student was enrolled in middle school, and the date (if applicable) in which a student left the sample, either due to school expulsion or moving to another school district. I note each out-of-school suspension, in chronological order, numbering them accordingly. I have also included information on each student's gender, race, and eligibility for free or reduced-price lunch.

Sample

I analyzed longitudinal data on the 1,693 students who comprised the cohort of students who enrolled in 6th grade in district middle schools during September and October of 2007. Note that in this study, I did not include students who moved *into* district middle schools (and into this cohort) after the first academic quarter of 6th grade. But, as is common in public schools, a few students in each grade arrived or left the district during each week of a school year. About 0.1% of the grade cohort left district

schools on a typical week during the school year. About 3.5% of the sample students left district schools during summer breaks after 6th grade and 7th grade. In June 2010, 1,460 students remained in the sample at the end of 8th grade. The other 14% of sample students had left the sample schools, and had likely moved to other school districts. I included all of the sampled students who began 6th grade in my analysis until they were *censored* upon their departure from the district.

The sample was split evenly between boys and girls. Nearly 45% of the students in the sample were eligible for free or reduced-price lunch. About half of the students were White, and about a quarter were Black. The remaining students were Asian, Hispanic or Native American. However, because only a very small number of students from these latter racial/ethnic categories were suspended repeatedly from school, I analyzed data on Black students and White students only (1,274 students), when disaggregating estimates of risk by race.

In total, the period from the first day of 6th grade until the last day of 8th grade covered 1,015 calendar days (from September 1, 2007 through June 11, 2010). But since a typical American child attends school for less than half of the days of the year, I secured archived district calendars identifying the dates of each school day for the years under study. I counted only the school days of an "academic year" when timing suspension and estimating its risk of occurrence. Thus, in this analysis, I measure time in the unit of *school days*. I recorded the dates of school enrollment, exit, and recorded disciplinary incidents using this measure of time, corresponding to three-years' of school days, during the entire span of 6th, 7th, and 8th grade. Thus, students were not officially at risk of out-of-school suspensions on weekends, holiday breaks, and summer recess.

Middle-school students in my sample could have attended 177 school days during 2007-08, 175 days during 2008-09, and 177 school days during 2009-10.

Additionally, students were not at risk of suspension when they were, in fact, already barred from attending school due to their behavior. The number of days associated with an out-of-school suspension for such students varied from 0.5 days (for a student suspended for the rest of the day, due to an infraction occurring *during* the school day) up to a maximum of five days. However, the median and the modal number of days of suspension was one day. Less than 10% of out-of-school suspensions imposed by school administrators were for more than three days.

Measures

I first organized the merged dataset in a multiple-record-per-subject format appropriate for survival analysis. Following Willett and Singer (1995), I divided every student's suspension history into spells. In each record, I recorded the finite span of time during which each student was enrolled and at risk of suspension, including the day of an incident triggering a school suspension, if and when it occurred. I label this chunk of time in which a student was first at risk of suspension the first *spell* of time. I labeled subsequent spells as the second spell and the third spell, as applicable. And each spell ended with either a student's suspension, or when the student's data were *censored* at the very end of a student's period at risk. At censoring, we do not know whether a student was suspended beyond the time that s/he was at risk. But once censoring occurred, students were no longer eligible to enter a subsequent spell.

In this analysis, I limit my investigation of the risk of suspension to the first three spells, as this exhausts much of the information in the dataset. Less than 7% of the

sample experienced more than three spells during their middle-school careers. And when including student demographic characteristics in the models, the event histories of students who experienced more than three suspensions were unsuitable for further modeling. Less than 5% of the female students and only 2% of the White students were at risk of a fourth suspension. Furthermore, the number of White female students who were at risk of a fourth suspension was negligible.

To clarify this formatting of spells, I present, in Table 1 and Figure 1, the histories of three prototypical students. In Panel A of Table 1, notice that multiple data lines correspond to time spans occurring during students' middle-school careers. The *Student* column identifies the individual student. *Spell* identifies each period of time during which a student is at risk for being suspended for the first, second, or third time. *Start Day* identifies the first day of a period of school days—typically the beginning of the school year or the beginning of a new *Spell* following a suspension. *End Day* marks the last day of each time period. *Suspend* functions as the censoring variable, indicating whether a student was suspended on the last day of the time period from *Start* to *End*. The next column records the number of days a student was suspended following a disciplinary incident. In Figure 1, I present a visual summary of the information from Table 1.

Table 1.

Panel A. Structure and contents of the multiple-record-per-subject survival dataset, for three prototypical students (n_{days}=1,509).

Record	Student	Spell	Start Day	End Day	Suspend	# Days Suspended
1	A	1	1	529	0	-
2	B	1	1	300	1	1
3	B	2	302	450	1	3
4	B	3	454	529	0	-
5	C	1	1	140	1	2
6	C	2	143	385	1	3
7	C	3	389	460	1	Expelled

Panel B. Structure and contents of the person-period survival dataset, for three prototypical students (n_{days}=1,509).

Record	Student	Spell	Day	Time	Suspend	# Days Suspended
1	A	1	1	1	0	-
2	A	1	2	2	0	-
:	:	:	:	:	:	:
528	A	1	528	528	0	-
529	A	1	529	529	0	-
530	B	1	1	1	0	-
531	B	1	2	2	0	-
:	:	:	:	:	:	:
828	B	1	299	299	0	-
829	B	1	300	300	1	1
830	B	2	302	1	0	-
831	B	2	303	2	0	-
:	:	:	:	:	:	:
978	B	2	450	149	1	3
979	B	3	454	1	0	-
980	B	3	455	2	0	-
:	:	:	:	:	:	:
1054	B	3	529	76	0	-
1055	C	1	1	1	0	-
1056	C	1	2	2	0	-
:	:	:	:	:	:	:
1194	C	1	140	140	1	2
1195	C	2	143	1	0	-
1196	C	2	144	2	0	-
:	:	:	:	:	:	:
1437	C	2	385	243	1	3
1438	C	3	389	1	0	-
1439	C	3	390	2	0	-
:	:	:	:	:	:	:
1509	C	3	460	72	1	Expelled

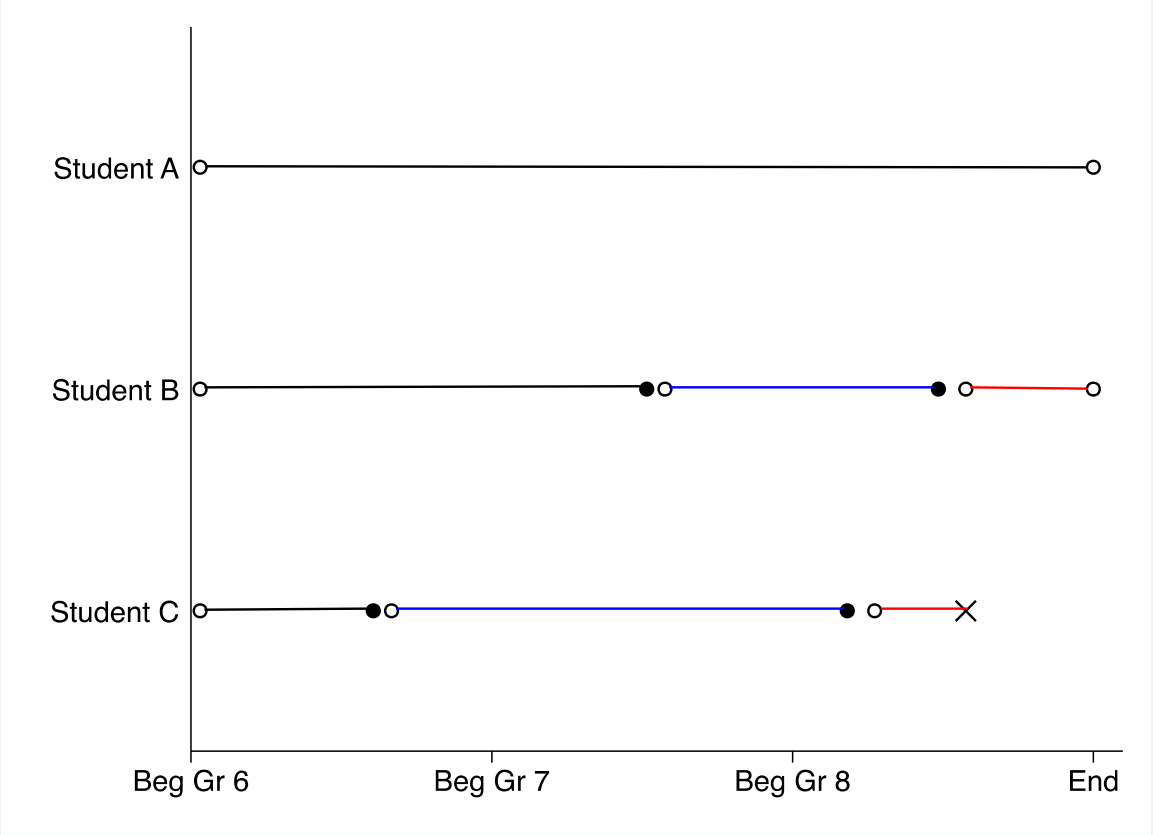

Figure 1. Graphical representation of the multiple-suspension history for three prototypical students, signifying three common configurations of the pattern of out-of-school suspensions.

Notice that the three students contributed different configurations of enrollment information and suspension histories. First consider Student A. Like most of the sample students, Student A entered the risk set at the beginning of 6th grade and completed 8th grade without being suspended. In Figure 1, Student A's suspension history is drawn as an uninterrupted straight line. And in Table 1, only the first record in the table is assigned to Student A. Thus, Student A spent his whole career in the first *SPELL*. He was censored on exit, because he left middle school without experiencing the event of interest, an out-of-school suspension.

Alternatively, Student B, beginning with record 2, enrolled at the beginning of 6th grade. She was suspended for the first time in her middle-school career—for one day—

during the spring of 7th grade, terminating record 2. In record 3, she returned to the risk set to start her second *SPELL*, on *DAY* 302. But midway through 8th grade she was suspended a second time, for three days, on *DAY* 450. She then returned to school for her third *SPELL* on *DAY* 454, at risk for a 3rd suspension that never occurred before she was *censored* at the end of 8th grade.

Finally, Student C is a hypothetical student who was suspended once in 6th grade and a second time during 8th grade. Then, approximately halfway through 8th grade, a 3rd incident—designated with an "X"—indicated that he was expelled. While a small number of middle-school students were expelled from the sample—and some of them were expelled for a limited time before returning to school—I chose to remove students from the sample following an expulsion, reasoning that their programming may have changed substantially upon returning to school. In the case of Student C, this incident also marks a third suspension.[18]

Before conducting my multiple-spell survival analyses, I stored the event-history data on the timing of suspensions for all students in a person-period data set (Singer & Willett, 1993). In Panel B, Table 1, I display this format for these same three prototypical students. In this dataset, each student contributes a record (row) for each school day that s/he was enrolled and at risk of being suspended. When students had been suspended, they were not at risk for a subsequent suspension until they returned to school after their

[18] Public-school principals are authorized only to *suspend* students, even for the most notorious rule breaking. Principals may—and in some cases are required to—*recommend* students for expulsion to the local Board of Education. However, I do not possess any qualitative information about the severity of any particular suspension incident. I only excluded students if the expulsion happened, and not if they did had been suspended following an event that perhaps warranted expulsion, but was not expelled for undisclosed reasons.

absence due to the disciplinary sanction imposed. Thus, for example, the histories of three prototypical students that I described above now are recorded in a person-period dataset with 1,509 records, corresponding to the 1,509 days in which the three students in question were at risk of suspension.

Student A, who completed middle school without being suspended, is represented by 529 records—all in *SPELL* #1. Student B is at risk for 525 school days, (and 525 records): 300 days at risk in *SPELL* #1, terminated on *DAY* 300 with a suspension; 149 days at risk in *SPELL* #2, terminated on *DAY* 450 with her second suspension; and 76 days at risk in *SPELL* #3, when she is censored at the end of 8th grade. And Student C is represented on 447 records: 140 days during *SPELL* #1, 210 days during *SPELL* #2, and 72 days during *SPELL* #3, before he was expelled—and therefore censored—on *DAY* 460, in the middle of 8th grade.

In addition to identifying each *DAY* that students were at risk, I also added a variable to count—beginning at *TIME* 1—the number of days students were present in each *SPELL*. Thus, for Student A, who was never suspended and therefore never left the first spell, *DAY* and *TIME* are equivalent in their values. For Student B *DAY* and *TIME* correspond for the first 300 records. Then, in the 301st row for Student B, *DAY* is recorded as 302 (recall that she was suspended during *DAY* 301), and *TIME* begins again at 1. Thus, in the person-period dataset, the last record terminating the second *SPELL* for Student B is marked as *DAY* 450 and *TIME* 148.

Outcome Variable:

In my analyses, my conceptual outcome is the risk of a suspension occurrence, represented by the *hazard probability*. It is defined as the conditional probability that

student *i* will be suspended from school on school day *t*, given that he or she was not suspended on any previous day during *SPELL j*. As such, this conditional risk probability is latent (not observed), but I do observe its concrete realization, defined as follows:

SUSPEND is a dichotomous variable, coded "1" if a student experienced a suspension from school, and "0" if the student was not suspended, on each school day that a student was at risk.

Note that, *within spell*, the coding of this outcome is intrinsically conditional. This is because, within the first spell, say, a student can only be coded "1" (suspended) providing s/he has not been suspended to that point. Once s/he is suspended (*SUSPEND*=1), the first spell is terminated. Upon returning to school, the student enters the second spell of attendance, and is again present (*SUSPEND*=0) until the next suspension, which will terminate the second spell, and so on. A student who is never suspended remains in the first spell for his or her entire career. All corresponding values of *SUSPEND* remain at a value of zero, and we say that the student is "censored at the end of the first spell," because they were never suspended while enrolled in district middle schools. A corresponding definition of censoring applies to all subsequent spells, should they occur.

Design Covariates:

SPELL is a time-varying integer variable that records the spell in which a student is at risk of suspension. It is coded "1" if a student is at risk for her/his first suspension from school, "2" if a student has been suspended once and is therefore at risk of a second suspension, and "3" if suspended twice and is at risk for a third suspension. In my dataset, it is recorded as a vector of dichotomous variables,

SPELLj, each coded "1" or "0" to identify each *SPELL.* I denote these three dummies by the vector **SPELL**, in the statistical models specified below.

TIME records the time, *t*, measured in school days, that have passed from the beginning of the respective *SPELL* until a student was suspended from school (if a student was indeed suspended) or censored (if they were not suspended to terminate this *SPELL*). In preliminary analyses, I tested different specifications for the continuous effect of time in my statistical models, combining both polynomial and logarithmic specifications.[19] I found that a polynomial logarithmic specification of time most suited this analysis. Consequently, I included the natural logarithm of predictor *TIME*, which I renamed *LNTIME*, and also the corresponding quadratic and cubic specifications, *LNTIME2* and *LNTIME3.* I denote the linear, quadric, and cubic functions of log-time by the vector **LNTIME**, in the statistical models specified below. Because of this use of the logarithmic time specification, and to avoid unnecessary infinities, I coded predictor *TIME* with the value 1, on the first day of each spell. Consequently, the value of *LNTIME* was 0 on that day.

Question Predictors:

MALE is a time-invariant dichotomous variable describing the gender of each student. It is coded "1" if a student is male, and "0" if a student is female.

BLACK is a time-invariant dichotomous variable describing whether a student self-identified as Black when enrolling at school. It is coded "1" if a student is Black, and "0" if a student is White. *When including a racial/ethnic designation in my*

[19] Results of the preliminary analyses are available from the author.

models, I remove Asian, Hispanic, and Native American students from that portion of the analysis, as there were too few of these latter students with enough suspensions to provide sufficient statistical power for credible comparisons among the groups. However, in models in which I do not account for race/ethnicity, I retain the full sample of students.

FRPL is a time-varying dichotomous variable describing whether a student was eligible to receive free or reduced-price lunch at school on each school day. It is coded "1" if the student was eligible for free or reduced-price lunch, and "0" otherwise. In my data *FRPL* is reported for the school year.

Note, students in my sample attended—and were therefore grouped within—different schools in the district. (Some students even moved *between* schools within the district during the period of observation.) To account for this, I included the random effects of school in my statistical models, distinguishing among schools using a SCHOOL ID variable.

Data-Analytic Plan

In my analyses, I hypothesize that on the first day of 6th grade—the beginning of their first "spell" of middle-school attendance—all students were at risk of a *first* out-of-school suspension from middle school. This risk—represented by the hazard function, $h_{i1}(t)$—describes the conditional probability that a student would be suspended on a particular school day, denoted by the value of the predictor $TIME_i$, in the first spell, given that they had not yet been suspended from middle school, earlier, as follows:

$$(1)\ \ h_{i1}(t) = \Pr\{TIME_i = t | TIME_{ij} \geq t, j = 1\}$$

Once suspended, a student then enters a second spell of attendance, in which s/he is again eligible—or "at risk"—for a further suspension. In general, the risk of the $(j+1)^{th}$ suspension began upon students' return to school after immediately serving the prescribed suspension days for the j^{th} incident:

$$(2)\ \ h_{ijt}(t) = \Pr\{TIME_{ij} = t | TIME_{ij} \geq t, j = 1,2,3\}$$

To address my research questions, I used logistic-regression analysis to fit discrete-time survival analysis (DTSA) models in the person-period dataset, to investigate the relationship between my dichotomous operationalization of the risk of suspension (*SUSPEND*) on sensible specifications of the spell and time predictors and a selection of substantive predictors. I conduct these analyses—simultaneously, across all three spells—in a multiple-spells discrete-time survival analysis (Willett & Singer, 1995). Thus, in general, I chose to model the risk of suspension—that is, the hazard probability, $h_{ijt}(t)$—as a function of spell, time, and individual student covariates, in logistic-regression models of the following generic form:

$$(3)\ \ h_{ijt}(t) = \frac{1}{1 + e^{-g(SPELL,TIME,COVARIATES)}}$$

Or, by taking natural logarithms and re-organizing:

$$(4)\ \ \text{logit}\ (\ h_{ijt}(t)) = g(SPELL, TIME, COVARIATES)$$

Below, I describe the various alternative specifications of this general model that I fitted to address each of my research questions.

RQ1: *When* are students at the greatest risk of being suspended—either once, or repeatedly—from public middle schools? Specifically,

a. **Within a spell, is the risk of suspension higher at particular time periods, lower at others?**

b. Does the temporally-dependent profile of risk differ by spell? Specifically, is the risk *higher* in subsequent spells than the risk of a first suspension, and if so, by how much? And, net of its level, does the shape of the risk profile differ by spell?

In addressing the first part of my first research question, I simply seek to summarize and describe the aggregate temporal profile of risk of suspension, by time-within-spell and by spell, over the entire middle-school career, regardless of the students' demographic characteristics. Consequently, I first fit the following general discrete-time survival analytic model:

$$(5)\ \text{logit}\left(h_{ijkt}(t)\right) = \boldsymbol{\beta_1}'\boldsymbol{SPELL_{ik}} + \boldsymbol{\beta_2}'\boldsymbol{LNTIME_{ik}} + \boldsymbol{\beta_3}'\boldsymbol{SPELL_{ik}} \times \boldsymbol{LNTIME_{ik}} + \zeta_k$$

where $h(t_{ijkt})$ describes the population conditional hazard probability—the *risk* of suspension—for student *i*, at risk during spell *j,* while attending school *k.* Parameter vector $\boldsymbol{\beta_1}$ describes the magnitude of the hazard, in aggregate, by spell. Parameter vector $\boldsymbol{\beta_2}$ describes the magnitude of the hazard, in aggregate, by time, and parameter vector $\boldsymbol{\beta_3}$ represents their interaction. I also include a random intercept, ζ_k, to represent differences in the risk of suspension associated with attending school *k* in all of my models.[20] During my analysis, via judicious comparisons of model fit between this general model and models that were more parsimonious, I eliminated unnecessary terms, leading to simpler

[20] In this case study, I found little evidence that students were grouped into only a few schools with high suspension rates. Nor were district schools substantially segregated by race or poverty—a potential bias of the effects of these covariates. In my subsequent Threats to Validity section, I address these concern by removing the random intercept from the models and comparing parallel parameter estimates. I also employed fixed-effects models as a robustness check for this same reason.

representations of the risk function within and between spells. This permitted me to address both parts of my research question simultaneously. To present my findings, I rely on the display of fitted risk functions for prototypical students, by spell.

RQ2: Over all spells, how does the profile of risk *differ* by key demographic dimensions?

In addressing this second research question, I focus on the impact on the risk of suspension of three important student characteristics: their gender, race/ethnicity, and eligibility for free or reduced/priced lunch. Student gender and race/ethnicity are, typically, visually salient, while the third dimension—a measure of student poverty—is less apparent. But students' eligibility for free or reduced-price lunch is readily available to school principals. (The school secretary likely collected the necessary forms to declare eligibility for this subsidy.) And all three are dimensions of student records necessary to report formal discipline to the state.

(a) Are middle-school boys at greater risk of suspension than middle-school girls? To continue the modeling process and include these critical dimensions, I first add to the statistical model in (4) above, the effect of *MALE* and any statistical interactions between *MALE* and *SPELL*, as follows:

$$
\begin{aligned}
(6)\ \text{logit}\left(h_{ijkt}(t)\right) &= \boldsymbol{\beta_1}'\boldsymbol{SPELL_{ik}} + \boldsymbol{\beta_2'LNTIME_{ik}} + \boldsymbol{\beta_3}'\boldsymbol{SPELL_{ik}} \times \boldsymbol{LNTIME_{ik}} \\
&+ \boldsymbol{\beta_4}'\boldsymbol{MALE_{ik}} + \boldsymbol{\beta_5}'\boldsymbol{SPELL_{ik}} \times \boldsymbol{MALE_{ik}} + \boldsymbol{\beta_6}'\boldsymbol{LNTIME_{ik}} \times \boldsymbol{MALE_{ik}} \\
&+ \boldsymbol{\beta_7}'\boldsymbol{SPELL_{ik}} \times \boldsymbol{LNTIME_{ik}} \times \boldsymbol{MALE_{ik}} + \zeta_k
\end{aligned}
$$

where parameter vectors $\boldsymbol{\beta_4}$, $\boldsymbol{\beta_5}$, $\boldsymbol{\beta_6}$and $\boldsymbol{\beta_7}$ represent the risk of being a male student, along with its statistical interactions between gender and both spell and time. Testing the

associated parameters allowed me to investigate whether the pattern of risk differed in *shape* as well as *level*, from spell to spell, by *MALE*. By inspection of model fit, I tested higher-order terms that were not required in the model and eliminated them.

(b) Are Black students at greater risk of suspension than White students? As noted earlier, for this sub-question (and for the final sub-question), due to limitations of my data, I continue only with a subsample of the cohort, eliminating Asian, Hispanic, and Native-American students from the risk set. I fitted generic multiple-spell discrete-time hazard models of the following form:

$$(6)\ \text{logit}\left(h_{ijkt}(t)\right) = \boldsymbol{\beta_1}'\boldsymbol{SPELL_{ik}} + \boldsymbol{\beta_2'LNTIME_{ik}} + \boldsymbol{\beta_3}'\boldsymbol{SPELL_{ik}} \times \boldsymbol{LNTIME_{ik}} + \boldsymbol{\beta_4}'\boldsymbol{BLACK_{ik}} + \boldsymbol{\beta_5}'\boldsymbol{SPELL_{ik}} \times \boldsymbol{BLACK_{ik}} + \boldsymbol{\beta_6}'\boldsymbol{LNTIME_{ik}} \times \boldsymbol{BLACK_{ik}} + \boldsymbol{\beta_7}'\boldsymbol{SPELL_{ik}} \times \boldsymbol{LNTIME_{ik}} \times \boldsymbol{BLACK_{ik}} + \zeta_k$$

where parameter vectors $\boldsymbol{\beta_4}$, $\boldsymbol{\beta_5}$, $\boldsymbol{\beta_6}$and $\boldsymbol{\beta_7}$ represent the risk of being a Black student, along with its statistical interactions with both spell and time. Testing the associated parameters allowed me to investigate whether the pattern of risk differed in *shape* as well as *level*, from spell to spell, by *BLACK*. By inspection of model fit, I tested higher-order terms that were not required in the model and eliminated them.

(c) Are students from low-income families at greater risk of suspension than students from more affluent families? For this sub-question, I return to the full dataset to explore differences in risk of suspension between students from low-income families and their more affluent peers by adding the effect of *FRPL*. I modeled the statistical interactions between *FRPL* and *SPELL*, as follows:

$$(7)\ \text{logit}\left(h_{ijkt}(t)\right) = \boldsymbol{\beta_1}'\boldsymbol{SPELL_{ik}} + \boldsymbol{\beta_2'LNTIME_{ik}} + \boldsymbol{\beta_3}'\boldsymbol{SPELL_{ik}} \times \boldsymbol{LNTIME_{ik}} + \boldsymbol{\beta_4}'\boldsymbol{FRPL_{ik}} + \boldsymbol{\beta_5}'\boldsymbol{SPELL_{ik}} \times \boldsymbol{FRPL_{ik}} + \boldsymbol{\beta_6}'\boldsymbol{LNTIME_{ik}} \times \boldsymbol{FRPL_{ik}} + \boldsymbol{\beta_7}'\boldsymbol{SPELL_{ik}} \times \boldsymbol{LNTIME_{ik}} \times \boldsymbol{FRPL_{ik}} + \zeta_k$$

where parameter vectors $\boldsymbol{\beta_4}$, $\boldsymbol{\beta_5}$, $\boldsymbol{\beta_6}$and $\boldsymbol{\beta_7}$ represent the risk of being a student eligible for free/reduced-price lunch, along with its statistical interactions with both spell and time. Testing the associated parameters allowed me to investigate whether the pattern of risk differed in *shape* as well as *level*, from spell to spell, by *FRPL*. By inspection of model fit, I tested higher-order terms that were not required in the model and eliminated them.

(d) Are the respective influences of gender, race, and poverty on the risk of suspension additive or interactive? For this final sub-question, again due to limitations of my data, I continued with a subsample of the cohort, eliminating Asian, Hispanic, and Native American students from the risk set. I modeled differences in risk of suspension for students by all three critical covariates—gender, race, and poverty—simultaneously, pooling together the models in (5) through (7) and including the higher-order interactions among the demographic characteristics of gender, race, and poverty and both spell and time. By inspection of model fit and model comparison, I tested higher-order terms that were not required and eliminated them.

Findings

The principals and assistant principals for the roughly ten schools[21] in my case study handled student discipline—including suspending students out of school—on most school days. Middle-school students in the district, as a group, were involved in incidents leading to an out-of-school suspension on more than 95% of the school days during the three academic years of this study. According to my analysis of the district academic calendar, middle-school students could have attended school on 529 days, over the three academic years from September 2007 until June 2010. And during this time, more than 4,500 different out-of-school suspension incidents occurred in the middle schools, an average of more than eight per day. On 85 school days (16%) there were three or fewer suspensions from district schools. But on 17 school days (3%), more than twenty incidents resulting in school suspensions had occurred.

In this study, I assume that *all* students in the sample were at risk of a school suspension, even if they "survived" middle school without a mark on their formal disciplinary record. On a typical school day, about two students from this sample were suspended. Sometimes none of the sampled students were suspended. And on two-thirds of the school days, two or fewer students from the sample were suspended. But conversely, on rare, busy days, seven or more of the sampled students were suspended on the same school day. Recall, also, that disciplinary incidents involving students in other grades had likely also occurred, drawing their principals' attention.

[21] I have deliberately obscured the exact number of schools and the relative size of each school to help protect the anonymity of district schools and students.

RQ1: *When* are students at the greatest risk of being suspended—either once, or repeatedly—from public middle schools?

In order to present and explore the risk of out-of-school suspension visually prior to reporting the results of my discrete-time survival analyses, I combine sample estimates of the survival probability until a first suspension (if it occurred) with a plot of subsequent suspensions for each student in the sample, in Figure 2. In the figure, to summarize the occurrence of the *first* suspension from middle school—or, alternatively, the distribution of students who survived middle school without being suspended—I superimpose a Kaplan-Meier (1958) estimate of the sample survivor function as a jagged black line that bisects the plot from top left to bottom right. This sample survival function describes the probability that a student had *not* been suspended (for the first time in middle school) prior to a given school day. Time, measured in school days, is denoted on the horizontal axis. On the vertical axis, I present the percentage of students who were at risk and had not yet been suspended, on each school day. Notice that, starting at the upper left corner, when all students have yet to be suspended from middle school, each *first* suspension that occurred (on one of the 529 different school days) nudges the sample survivor curve downward. The suspended student was then eliminated from the risk pool for the first *SPELL*, with nearly all of these having then joined the second *SPELL*. Because about 80% of the sample students "survived" without ever being suspended from middle school, the bottom of the truncated plot, from 0% through 80%, then, is completely blank.

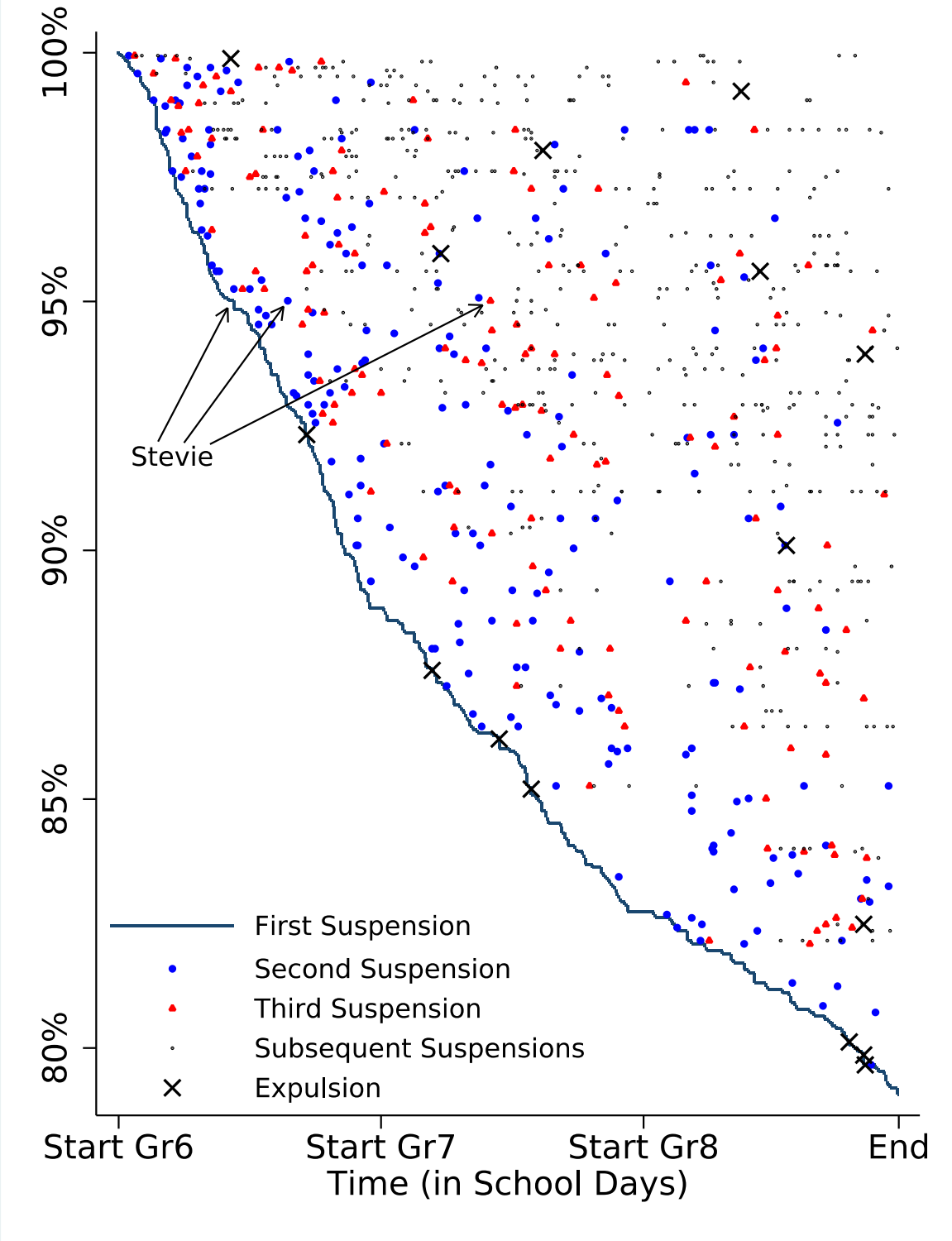

Figure 2. Kaplan-Meier survivor plot of students' first suspension from middle school, combined with an abacus plot of students' second, third, and subsequent suspensions, if required, and describing one example student ($n_{students}$=1,693, $n_{1stSuspension}$=340, $n_{2ndSuspension}$=210, $n_{3rdSuspension}$=139, $n_{SubsequentSuspensions}$=494).

Note, also, that the plotted black survival curve denotes, approximately, the school day marking the beginning of students' second *SPELL*. (The small gap of time that represents the duration of any suspension—five days or fewer—when suspended students were not at risk is not shown on this plot.) Note that 333 students were suspended from middle school for the first time, and were readmitted into school following this first suspension. Seven students were expelled from school at their first suspension—these students are denoted by an "X" on the plot—and, by virtue of their expulsion, they left the risk set for all subsequent *SPELL*s on their first suspension.

While the probability of being suspended on any given day may seem large to a young student who is worried about the *threat* of being suspended at all, in reality this risk of a first suspension on any *particular* school day was actually quite small. This low daily probability of being suspended for the first time is a function of the large number of days that the 1,693 students were actually at risk of a first suspension: a total of 735,113 person-school days. Still, as risk accumulated over time—during the course of the 529 days of middle-school attendance—eventually about 20% of the sample students (340) were suspended from middle school at least once. Conversely, 1,353 students were never suspended from middle school, before they either completed 8th grade or, perhaps, transferred to another school district and were therefore *censored* from the risk set.

I display the incidence of a second suspension (if it occurred) as a blue circle aligned horizontally (like a bean on an abacus) with their location on the first-spell survivor plot to a position to its right, marking the end of the second *SPELL*. For example, one student, labeled "Stevie" in the figure, was first suspended in the middle of 6th grade (where the estimated survivor function intersected the 95% survival-probability

level). He was readmitted to school following this first suspension, but was suspended again approximately 35 school days later—while still in 6th grade—as denoted by the blue circle directly, and horizontally, to the right of when he was first suspended. Continuing to follow "Stevie's" progress through his third *SPELL*, he was suspended again near the middle of 7th grade. He and the other students who were suspended a third time are marked on the plot with a red triangle. "Stevie's" third suspension is marked along the same horizontal axis as his second suspension, and to its right. I show any subsequent suspensions as small dots on the plot.

Notice that fewer than 7% of the sampled students account for 476 subsequent suspensions, in *SPELL* #4 through *SPELL* #23. The small group of students who experienced multiple suspensions did not provide sufficient information (statistical power) to fit discrete-time hazard models, and estimate the risk of suspension precisely beyond the third spell, particularly when the risk set is partitioned by gender, race and poverty. Thus, in the discrete-time hazard analyses that I present here, I have focused on only the occurrence of the first three instances of suspension from middle school.

To address my first research question, estimating these three risk profiles, I fitted the discrete-time hazard model specified in Equation 5—and several subsequent "reduced" models of increasing parsimony—in the person-period dataset. In Table 2, I present the results of fitting three such models, labeled as Models A, B, and C, each of decreasing complexity (increasing parsimony) than the one that precedes it.

Table 2. Parameter estimates, approximate *p*-values, and goodness-of-fit statistics from three fitted multiple-spell hazard models predicting the risk of out-of-school suspension by both spell and time within spell ($n_{students}$=1,693, $n_{suspensions}$=691).

	Model A	Model B	Model C
SPELL1	-10.196***	-10.324***	-6.804***
SPELL2	-5.198***	-4.860***	-4.651***
SPELL3	-4.763***	-4.925***	-4.098***
LNTIME×SP1	1.483	1.590***	-0.183***
LNTIME×SP2	0.463	-0.061	-0.212***
LNTIME×SP3	0.112	0.377	-0.264***
LNTIME2×SP1	-0.178	-0.205***	
LNTIME2×SP2	-0.214	-0.022	
LNTIME2×SP3	0.002	-0.101*	
LNTIME3×SP1	-0.002		
LNTIME3×SP2	0.02		
LNTIME3×SP3	-0.011		
rho	0.045	0.045	0.045
rank	13	10	7
-2LL	10076.8	10078.0	10115.8
Likelihood -Test (compared to Model A)		1.2	37.815
df		3	3
p-value		*0.753*	*<0.001*

*~ p<0.10, * p<0.05, ** p<0.01, *** p<0.001*

In the table, I first list parameter estimates and goodness-of-fit statistics for a discrete-time hazard model with a fully interacted specification of the time predictors *SPELL* and *LNTIME*, with the latter exhibiting linear, quadratic and cubic components (Model A). Noticing that only estimates of the "main effect" of *SPELL* were large and statistically significant in this fitted model, I dropped the highest-order interaction terms—the two-way interactions of *SPELL* and cubic *LNTIME*—from the model and thereby fitted reduced Model B, containing linear and quadratic specifications of the interaction of spell and time, and the main effect of spell. Note that trimming these terms did not result in a statistically significant decrement to the fit of the model (χ^2=1.2, *df*=3,

p=0.753). Finally, I also fitted a third—and further reduced—model (Model C), in which I retained only the interactions of spell and linear *LNTIME*, in addition to the main effect of *SPELL.* However, I found that removing the interactions of spell and the quadratic specifications of time *did* result in a significant decrement in the fit of the model (χ^2=37.8, df=3, p=<.001). Therefore, I concluded that the model containing the main effects of spell and the interactions between spell and a quadratic specification of (log) time—Model B—best described the relationship among risk, spell, and time, in my data.

Even a cursory inspection of estimated parameters, in fitted Model B of Table 2, reveals that the risk of suspension by time differs dramatically between first and subsequent spells. Note, for instance, that the slope parameters associated with both linear and quadratic time in Spell #1—represented as two-way interaction terms in the fitted model—differ considerably between Spell #1 and other spells. However, the exact dependencies of fitted risk on both time and spell are complex. Thus, in Figure 3, to aid interpretation of the estimated risk profiles by spell and time in Model B of Table 2, I present a plot of the corresponding fitted hazard (top panel) and fitted survival (bottom panel) functions, obtained from fitted Model B.

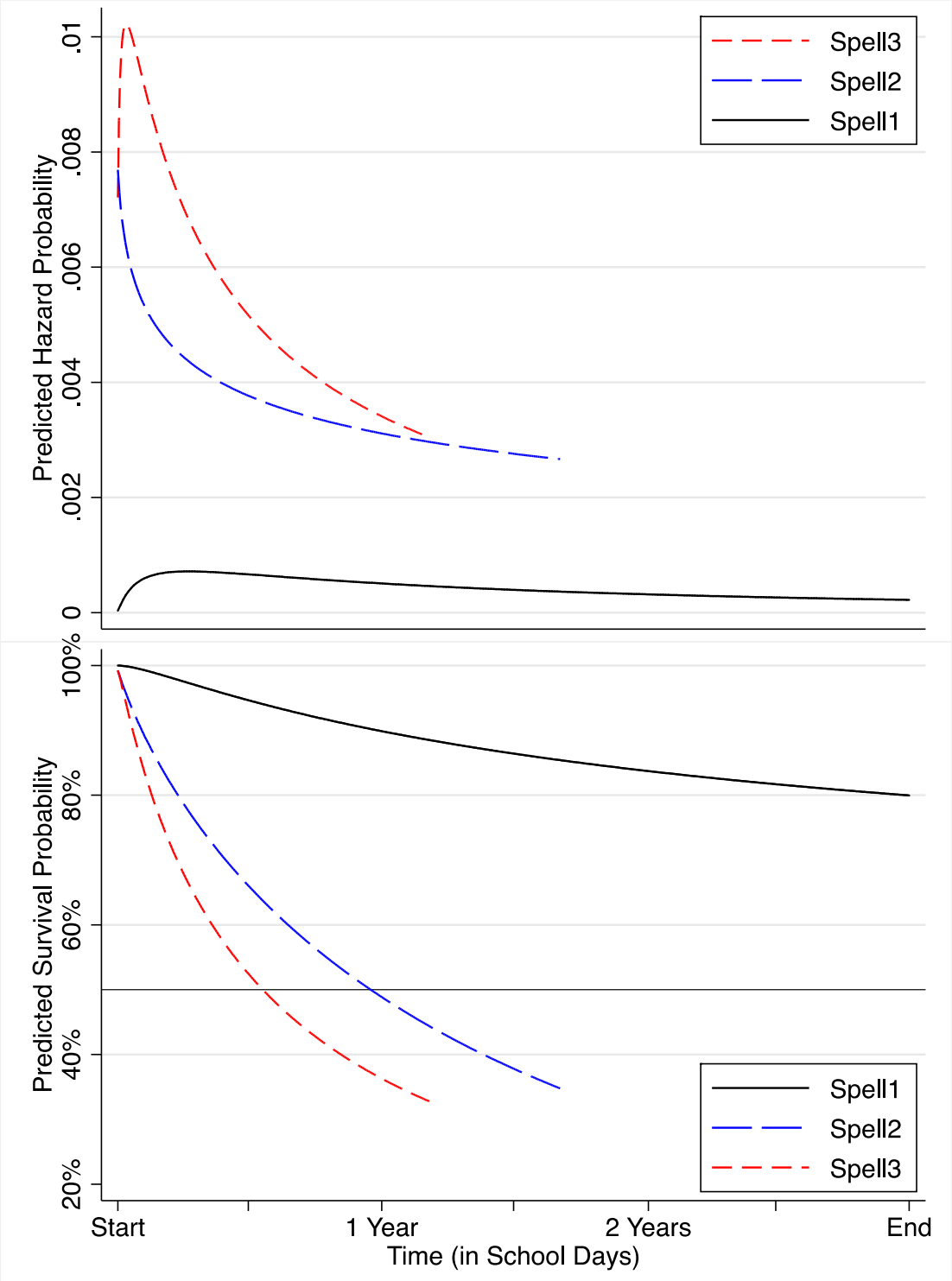

Figure 3. Predicted hazard probability (top panel) and survival probability (bottom panel) of a students' first, second and third suspension from middle school ($n_{students}$=1,693, $n_{suspensions}$=689). Predicted values obtained from fitted Model B in Table 2.

First, in the upper panel of the figure, notice that the risk of suspension in Spell #1 remained relatively small—less than one-tenth of one percent—over the entire middle-school career, when compared to the risk of a second or third suspension, if such subsequent suspensions even occurred. In addition, notice that the risk of a first suspension peaked early in 6th grade, and then slowly diminished throughout the rest of middle school. At its highest, I estimate the daily risk of first suspension was 0.00071 on school day 48—around the middle of November during 6th grade. By the end of 8th grade, however, the risk of a first suspension (for those students who had "survived" more than 500 days of middle school without being suspended) was approximately 0.00022 per day.

The conditional risk set for a *second* school suspension then consisted of the students who had been suspended once, and had then been re-admitted after their first suspension. As noted above, seven students were expelled from school in conjunction with their first suspension, leaving a risk set of 333 students to begin their second *SPELL*. And more than 60% of those eligible (210 of the 333 students) were indeed suspended a second time. Two students were expelled at the ends of their respective second *SPELL*s.

Notice that the fitted profile of risk of a second suspension, shown in blue in the upper panel of the figure, is *substantially higher* than the risk of a first suspension and has a considerably different profile. (Indeed, correspondingly, in Figure 2, note that a substantial number of the blue "dots" are clustered close to the date of students' first suspension, if they occurred at all.) From fitted Model B, in Table 2, I estimate that the risk of a second suspension was highest immediately upon readmission to school following a first school suspension. Note that, in Figure 3, the dashed blue line describing students' risk of a second suspension started well above 0.006. Even one year after a first

suspension, I estimate that the risk of a second suspension remained about 0.003. In the lower panel of the figure, I display the model-estimated survival probability.[22]

Similarly, the risk set for the third *SPELL* consisted of the students who were readmitted following their second suspension. About two-thirds of these students were then suspended a third time (139 of the 208 students). In shape, and again from Model B of Table 2, the fitted third-spell risk function is similar to the risk function for the second spell, but slightly elevated above it, indicating a greater risk of suspension for those who were suspended two times and then returned to school. In Figure 3, the dotted red line describing students' risk of a third suspension peaked earlier than the corresponding curves for the first two spells, and above 0.01. I estimate that the risk of a third suspension was highest on the sixth day upon readmission to school following a second school suspension. In layman's terms, students who had been suspended once, at any time during their middle-school career, were much more likely to be suspended again, and pretty soon after returning to school following their earlier suspension. This pattern is even more exacerbated between the second and third spells, than between the first and second.

In the bottom panel of Figure 3, I display corresponding model-predicted survival functions for these middle-school students. The fitted survival plot for Spell #1 corresponds well with the Kaplan-Meier-estimated plot that I provided as part of Figure 2, with 80% of the sample students "surviving" without being suspended before the end of 8th grade. For students who were suspended a first time, the estimated survival

[22] In display these plots for 95% of each risk set, truncating these profiles once the polynomial risk profiles are too small to be modeled meaningfully.

function for a second suspension is substantially shorter. Superimposed on the survival plot, I provide a reference line denoting the 50% survival mark. Projections onto the horizontal (time) axis then provide estimates of the median "lifetimes" in the second and third spell. These describe the time (in school days) that must past—on average—until half of the students who had been suspended a first time were suspended a second time, and then a third time. Thus, reading directly from the abscissa, I note that after 170 school days (about one year after a first suspension), I estimate that half of these students had been suspended a second time. And for those who were suspended a second time, the model-estimated survival time was shorter still. After 98 school days, half of the students at risk of a third suspension had indeed been suspended again. In my view as a former principal and educator, these are dramatic differences in the expression of the risk of suspension, by spell.

RQ2: How does the profile of risk *differ* by key demographic dimensions?

Mirroring national trends in school suspensions, simple suspension rates estimated empirically in my sample of 1,693 middle-school students (that is, estimated as sample statistics, without the benefit of discrete-time hazard modeling) also differ dramatically by both gender and race. For instance, sample statistics indicate that about 25% of the sampled boys were suspended, compared to only 15% of the girls. And while the suspension rate for White students (8%) in the sample was relatively low, more than half of Black students (55%) were suspended, at least once, during middle school. Additionally, in my sample, the suspension rate for students eligible for free or reduced-price lunch (36%) was more than four times the rate students not receiving this subsidy (8%). These differences in sample statistics by selected student demographic

characteristics provide the logical basis for reporting the findings of the discrete-time hazard analyses to follow, to address the several parts of my second research question, beginning with an exploration of differences in the risk of suspension by each of three important characteristics—student gender, race, and poverty—separately, and then examining their joint effects.

(a) Are middle-school boys at greater risk of suspension than middle-school girls? To estimate differences in the risk of suspension by student gender, I first extended Model B of Table 2 to specify the population risk of suspension as a function of: *SPELL*; linear, quadratic, and cubic *LNTIME*; the predictor *MALE* to distinguish students by their respective genders; the two-way interactions between *SPELL* and *LNTIME* (again with linear, quadratic and cubic specifications); the two-way interaction of *SPELL* with *MALE*; the two-way interaction of *MALE* with *LNTIME*; and the three-way interaction between *SPELL*, *LNTIME*, and *MALE*. This "full" model included 25 parameters representing the corresponding hypothesized fixed effects (and included one additional parameter to represent the random effect of school). Through judicious trimming of early models, based on comparisons of model fit among nested models, I first trimmed away the three-way interaction, then the three terms in cubic *LNTIME* (for the three *SPELL*s), then the two-way interaction of *MALE* with *LNTIME*—producing a much more parsimonious model with 12 fewer parameters, while yielding only a deviance statistic of 14 points, for a saving of 12 degrees of freedom (p=0.30). I also found that the difference in risk between boys and girls was concentrated entirely in the first *SPELL*, and consequently trimmed the two statistically non-significant parameters representing the effects of the $MALE \times SP2$ and $MALE \times SP3$ interactions in the model. And so, for a

final "reduced" model that predicts the risk of suspension for the first three *SPELL*s, by student gender, I specified and fit the most parsimonious model, Model D, whose parameter estimates and fit statistics I list in the second column of Table 3. Note that the chosen model—Model D—retained all the temporal features of the model that I used to address my first research question, and added the statistically significant effect of student gender in the first spell (χ^2=29.7, *df*=1, *p*<.001).

Thus, in Table 3, I present parameter estimates, approximate *p*-values, and goodness-of-fit statistics for Model D. (I also present similar statistics for three other final discrete-time multiple-spell hazard models, corresponding to research questions 2(b), 2(c), and 2(d) respectively, in which the risk of out-of-school suspension is distinguished by student gender, race/ethnicity, free/reduced-price lunch status, and the joint effect of all three covariates.)[23] From the estimated slope parameter associated with the $MALE \times SP1$ interaction in Model D (0.61, *p*<.001), I find that the fitted risk of a first suspension for a middle-school boy considerably greater than the corresponding risk for a girl. Specifically, anti-logging the estimate for this slope parameter, I conclude that the fitted odds of a first suspension for boys are 1.84 times the odds of a first suspension for girls. Moreover, because interactions between gender and the predictors distinguishing the second third spells were not retained in the parsimonious fitted model, I conclude that there were no gender-related differences in the risk of suspension beyond the first spell. In other words, once a middle-school student has been suspended a first time, the odds of subsequent suspensions were the same for boys and girls.

[23] I provide a full set of the models specified and fitted en route to my preferred parsimonious models in the Appendix.

Table 3. Parameter estimates, approximate *p*-values, and goodness-of-fit statistics from four fitted multiple-spell hazard models predicting the risk of out-of-school suspension by spell, time within spell, and critical covariates.

	Model D	Model E	Model F	Model G
SPELL1	-10.68***	-11.08***	-11.27***	-12.55***
SPELL2	-4.864***	-5.074***	-4.844***	-5.084***
SPELL3	-4.929***	-5.076***	-4.906***	-5.079***
LNTIME×SP1	1.590***	1.290**	1.485***	1.225**
LNTIME×SP2	-0.061	-0.142	-0.0559	-0.141
LNTIME×SP3	0.377	0.277	0.378	0.276
LNTIME2×SP1	-0.204***	-0.159***	-0.187***	-0.146**
LNTIME2×SP2	-0.0221	-0.00933	-0.0233	-0.00956
LNTIME2×SP3	-0.101*	-0.0856~	-0.102*	-0.0854~
MALE×SP1	0.612***			1.501***
BLACK×SP1		2.331***		2.195***
BLACK×SP2		0.598**		0.601**
BLACK×SP3		0.464~		0.461~
FRPL×SP1			1.710***	2.423***
MALE×FRPL×SP1				-0.949**
BLACK×FRPL×SP1				-0.929**
N	798,960	596,789	798,960	596,789
rho	0.047	0.022	0.033	0.024
rank	11	13	11	17
-2LL	10047.2	8412.0	9871.8	8305.2

*~ p<0.10, * p<0.05, ** p<0.01, *** p<0.001*

To illustrate these findings, in Figure 4, I present a plot of fitted hazard function (top panel) and fitted survival function (bottom panel), by student gender, with plotted values being obtained from fitted Model D. Again, in terms of net elevation, notice that the risk of suspension in Spell #1 remained relatively small compared to the risk of a second or third suspension, if such events occurred. I estimate that the risk of a first suspension peaked on school day 49, and then slowly diminished throughout the rest of middle school. However, the risk of a first suspension for girls on that day was 0.00051, while the risk of a first suspension for boys was 0.00093. Based on these fitted

probabilities, the odds of a first suspension for boys were 84% higher than the odds of a first suspension for girls, throughout middle school.

The conditional risk set for a *second* school suspension consisted of the students who had been suspended once, and had then been re-admitted after their first suspension. Again, as noted in my findings for the first research question, the profile of risk for a second suspension, shown in blue, is substantially higher than the risk of a first suspension. But this risk is independent of whether a student is a boy or a girl, because of the absence of the corresponding interaction term in the fitted model. Indeed, the risk profile for a second suspension for both boys and girls displayed in Figure 4 is virtually indistinguishable from the parallel risk profile in Figure 3, as one might expect, as all boys and girls are effectively pooled into the second- and third-spell risk estimates in Figure 4. Similarly, the profile of risk for a third suspension, shown in red, is also substantially higher than the risk of a first suspension, is also highest within the first few days upon readmission to school following a second school suspension, and likewise, the risk of a third suspension does not differ by student gender.

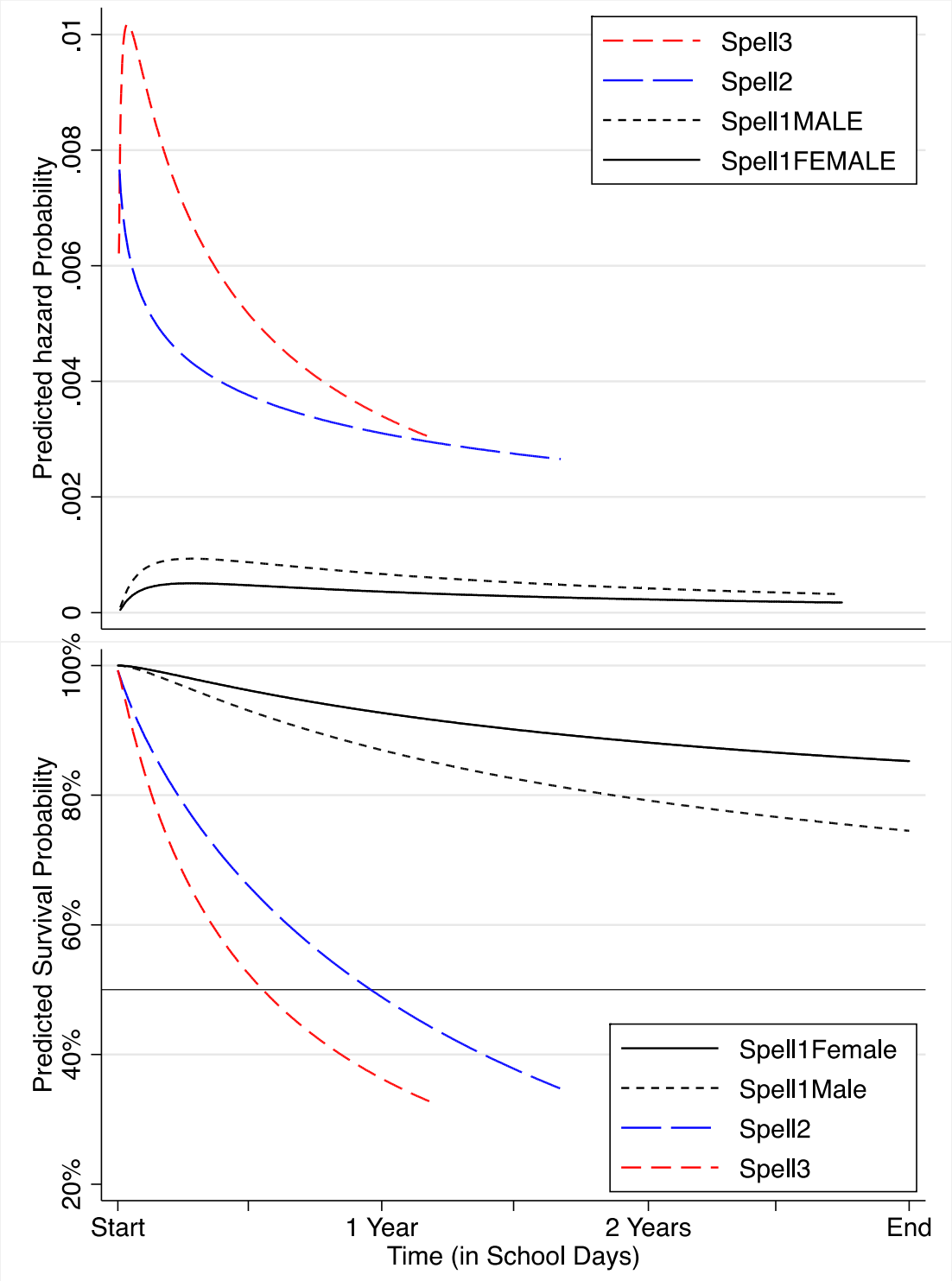

Figure 4. Predicted hazard probability (top panel) and survival probability (bottom panel) of a students' first, second and third suspension from middle school, disaggregated by student gender ($n_{students}$=1,693, $n_{suspensions}$=689). Predicted values obtained from fitted Model D in Table 3.

In the bottom panel of Figure 4, I display the corresponding fitted survival functions for these middle-school boys and girls. The fitted survival plot for Spell #1 shows that 75% of the boys and 85% of the girls "survived" without being suspended before the end of 8th grade. But once a student was suspended a first time, time to subsequent suspensions was measurably lower. In the plot, I provide a horizontal reference line denoting the 50% survival time (median lifetime), summarizing the time that must pass before half of the students who had been suspended previously were suspended again. In fitted Model D, after 170 school days (just less than one year after a first suspension) half of these boys and girls had been suspended a second time. Similarly, following a second suspension, I estimate that half of the students at risk of a third suspension were suspended within 98 school days.

(b) Are Black students at greater risk of suspension than White students? Following the same logical procedure as above, but estimating differences in the risk of suspension between Black students and White students, I first fitted a discrete-time hazard model in which I specified the population risk of suspension as a function of: *SPELL*; linear, quadratic, and cubic *LNTIME*; the predictor *BLACK* to distinguish students by their reported racial category; the two-way interactions between *SPELL* and *LNTIME* (again with linear, quadratic and cubic specifications); the interaction of *SPELL* with *BLACK*; the interaction of *BLACK* with *LNTIME*; and the three-way interaction between *SPELL*, *LNTIME*, and *BLACK*. This "full" model included 25 parameters (including one to represent the random effect of school). Again, through judicious trimming of the models, based on comparisons of model fit among nested models, I trimmed away the three-way interaction, then the three cubic terms (for the three

*SPELL*s), and then interaction of *BLACK* with *LNTIME*. Thus, I trimmed 12 parameters while yielding only a deviance statistic of 18.7 for a saving of 12 degrees of freedom (p=0.095).

Then, I tested whether the difference in risk between Black students and White students was concentrated in the first *SPELL*. However, I concluded that the differences in risk between Black and White students remained sizeable for subsequent suspensions. So, for my final fitted model in which I predict the risk of suspension for the first three *SPELL*s, by student race, I specified a "reduced" parsimonious model—Model E—whose parameter estimates are listed in the third column of Table 3.

From the fitted slope parameter associated with the $BLACK \times SP1$ interaction in Model E (2.33, p<.001), in the table, I estimate that the risk of a first suspension for a middle-school Black student is considerably greater than the risk for a corresponding White student. Specifically, anti-logging the estimate for this slope parameter, the fitted odds of a first suspension for Black students are *10 times* the odds of a first suspension for White students. Moreover, once a middle-school student was suspended a first time, I find that the risk of a second suspension for Black students, as estimated by the slope parameter associated with the $BLACK \times SP2$ interaction in Model E (0.598, p<.01), is again higher than the risk for a corresponding White student. Anti-logging this slope parameter, I conclude that the fitted odds of a second suspension for Black students were 82% higher than the odds for White students, *controlling for a first suspension.* Similarly, anti-logging the parameter estimating $BLACK \times SP3$ (0.464, p<0.10), I conclude that the fitted odds of a third suspension for Black students were 58% higher than the odds for White students, *controlling for a second suspension.*

To illustrate these findings, in Figure 5, I present a plot of the corresponding fitted hazard function (top panel) and the fitted survival functions (bottom panel), by race, with predicted values obtained from fitted Model E. First, notice that the risk of suspension in Spell #1—miniscule for White students and approximately 10 times that for Black students—peaked at 0.0219 for Black students and 0.0021 for White students, on school day 58 of 6th grade. Based on these fitted probabilities, the odds of a first suspension for Black students were ten times higher than the odds of a first suspension for White students, throughout middle school.

Again, the profile of risk for a second suspension, shown in blue, is substantially higher than the risk of a first suspension. But, now, this risk of a second suspension *also* depends on whether a student is Black or White—with the risk of a second suspension for White students quite high immediately after a first suspension, but tapering down quickly below 0.004. In contrast, the risk of a second suspension for Black students remains 82% higher, on each day of Spell #2. I estimate that the risk of a second suspension for Black students is higher than 0.004, even one year after returning from a first suspension.

Similarly, the profile of risk for a third suspension, shown in red, is also substantially higher than the risk of a second suspension, almost immediately after the beginning of the third *SPELL*. But, again, notice that the model-estimated risk of a third suspension for White students is lower than the risk of a second suspension for Black students, on each day following a previous suspension. Indeed, on all days of middle school, the model-estimated risk of a second suspension for a Black student is *higher* than the comparable risk of suspension for a White student who had already been suspended twice.

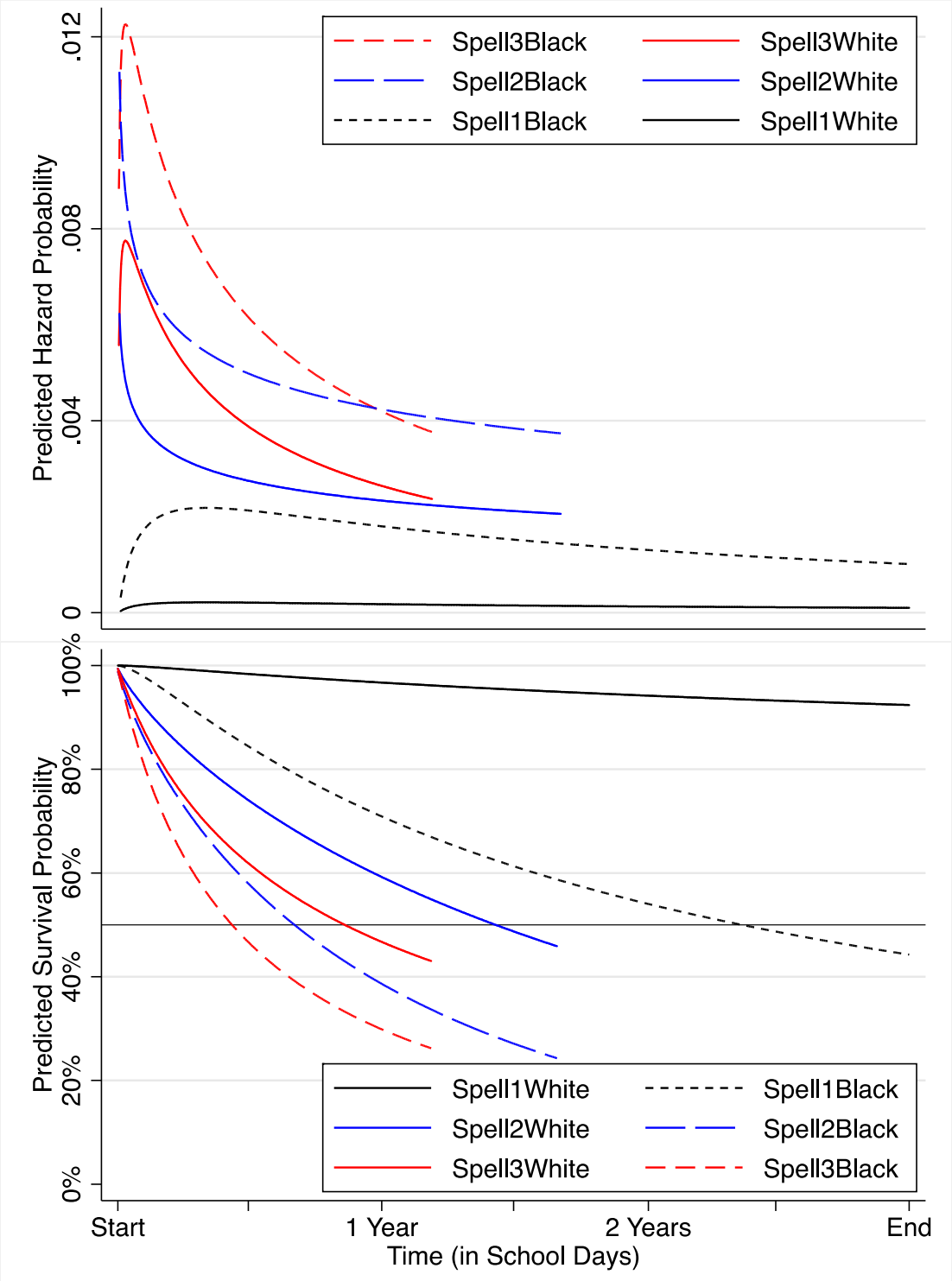

Figure 5. Predicted hazard probability (top panel) and survival probability (bottom panel) of a students' first, second and third suspension from middle school, disaggregated by student gender ($n_{students}$=1,274, $n_{suspensions}$=613). Predicted values obtained from fitted Model E in Table 3.

In the bottom panel of Figure 5, I provide a corresponding fitted survival plot for Model E. This plot is the compilation of six different estimates of risk—the risk of suspension for the first three spells for Black students and White students—over time. Note, first, that the vast majority—92%—of the White students "survived" middle school without being suspended, as denoted by the solid black line. In contrast, the model-predicted survival plot for Black students portrays a completely different manifestation of the risk. After approximately 418 school days (somewhere around Thanksgiving break of students' 8th grade year) half of the Black students in the sample had been suspended from middle school. Similarly, I predict that the median lifetime of a White student who had been suspended a first time in middle school "survived" approximately 253 school days until a second suspension (should it occur). In contrast, a Black student who had been suspended a first time "survived" only 119 school days until a second suspension from middle school occurred. Furthermore, I predict that the median lifetime of a White student who had been suspended a second time "survived" approximately 153 school days, while a Black student who had been suspended a second time "survived" for only 78 school days before being suspended a third time.

(c) Are students from low-income families at greater risk of suspension than students from more affluent families? Again, following the same logical procedure as above, but estimating differences in the risk of suspension between students eligible for free or reduced-price lunch and students who are not eligible for this subsidy, I first fitted a "full" discrete-time hazard model in which I specified the population risk of suspension as a function of: *SPELL*; linear, quadratic, and cubic *LNTIME*; the predictor *FRPL*; the two-way interactions between *SPELL* and *LNTIME* (again with linear, quadratic and

cubic specifications); the interaction of *SPELL* with *FRPL*; the interaction of *FRPL* with *LNTIME*; and the three-way interaction between *SPELL*, *LNTIME*, and *FRPL*. I again trimmed away the three-way interaction, the three cubic terms, and the interaction of *FRPL* with *LNTIME*. I found that the difference in risk between those eligible for free or reduced-price lunch and those students not eligible was concentrated entirely in the first *SPELL*. Consequently, I trimmed the two statistically non-significant parameters for the $FRPL \times SP2$ and $FRPL \times SP3$ interactions in the model, producing a parsimonious "reduced" model with 14 fewer parameters, while yielding deviance statistic of 18.0, for the saving of 14 degrees of freedom (p=0.21). Thus, in my final model predicting the risk of suspension for the first three *SPELL*s, by *FRPL*, I specified and fit the more parsimonious model—Model F—whose parameter estimates are listed in the fourth column of Table 3.

From the estimate of the slope parameter associated with the $FRPL \times SP1$ interaction in Model F (1.71, p<.001), I find that the risk of a first suspension for a student eligible for free or reduced-price lunch is substantially greater than the risk for a corresponding middle-school student who is not eligible for the subsidy. Specifically, anti-logging the estimate for this slope parameter, I conclude that the fitted odds of a first suspension for students eligible for free or reduced-price lunch are 5.5 times the odds of a first suspension for students not eligible for free or reduced-price lunch. And, paralleling the results for Model D, in which I explored the effect of student gender, once a middle-school student was suspended a first time, the fitted odds of subsequent suspensions were the same for all students who had been suspended previously, regardless of their eligibility for free or reduced-price lunch.

As before, in Figure 6, I present a plot of the fitted hazard (top panel) and survival (bottom panel) functions, by student eligibility for free or reduced-price lunch, with predicted values obtained from fitted Model F. Again, notice that the risk of suspension in Spell #1 remained relatively small, compared to the risk of a second or third suspension, if such events occurred. I estimate that the risk of first suspension peaked on school day 53, and then slowly diminished throughout the rest of middle school. The risk of a first suspension for students eligible for *FRPL* on that day was 0.00134. For students not eligible, the risk of a first suspension was 0.00024 on school day 53. Based on these fitted probabilities, the odds of a first suspension were 5.5 times for students receiving *FRPL* than the odds of a first suspension for those not receiving *FRPL*, throughout middle school.

As above, the conditional risk set for a second school suspension consists of students who had been suspended once and had then been re-admitted after their first suspension. The risk of a second suspension, shown in blue, is again substantially higher than the risk of a first suspension, but is independent of students' *FRPL* status, because of the absence of the corresponding interaction term in the fitted model. Similarly, the profile of risk for a third suspension, shown in red, is also substantially higher than the risk of a first suspension. And again, this risk does not differ by *FRPL* status. Thus, the risk profiles for a second suspension and third suspension (should they occur) are virtually indistinguishable in Figures 3, 4 and 6, as all of the students are effectively pooled into the second- and third-spell risk estimates.

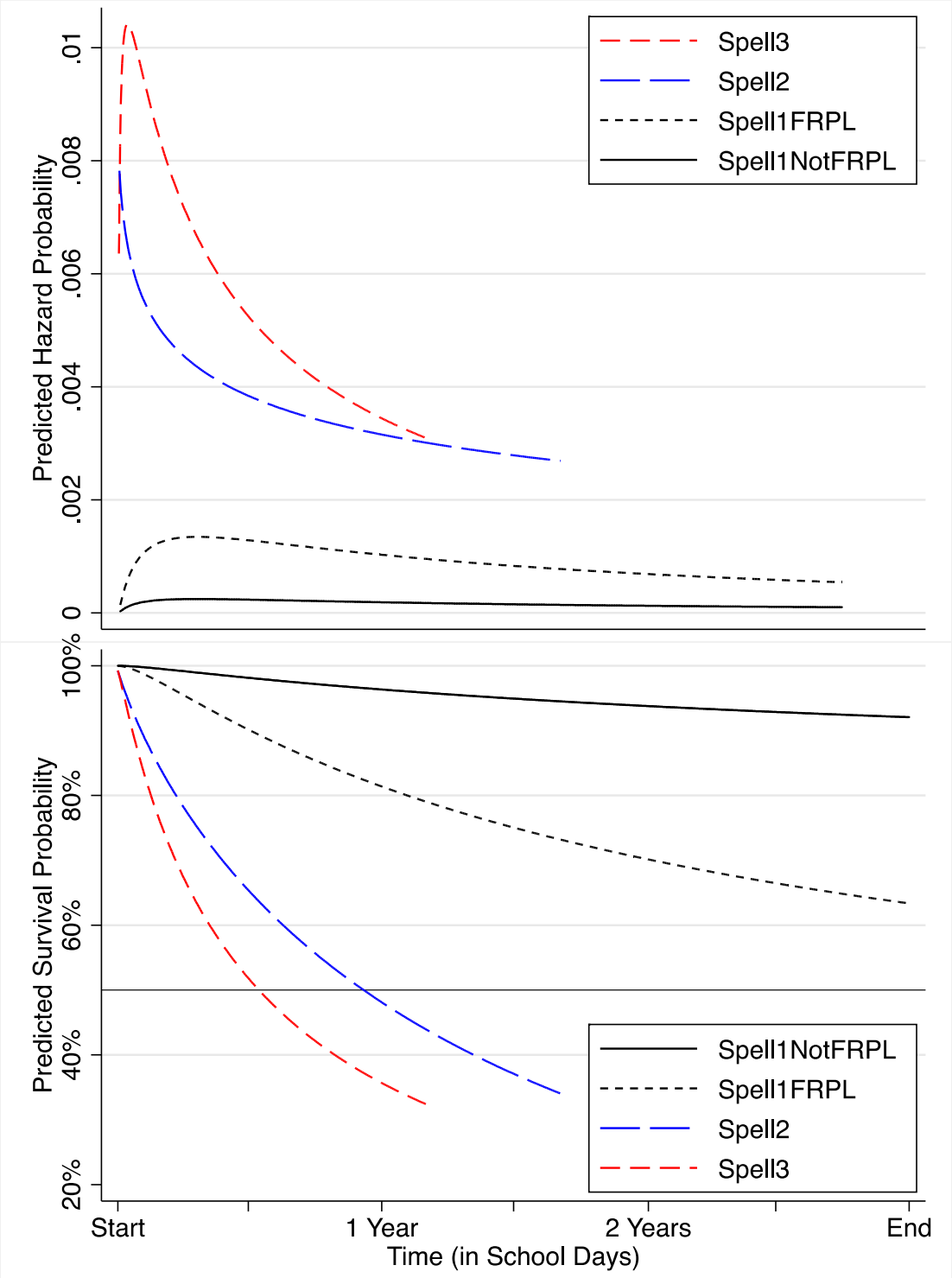

Figure 6. Predicted hazard probability (top panel) and survival probability (bottom panel) of a students' first, second and third suspension from middle school, disaggregated by free/reduced-price lunch status ($n_{students}$=1,693, $n_{suspensions}$=689). Predicted values obtained from fitted Model F in Table 3.

In the bottom panel of Figure 6, I provide a corresponding fitted survival plot for Model F. The fitted survival plot for Spell #1 shows that 63% of the students receiving *FRPL* and 92% of those not receiving *FRPL* "survived" without being suspended during middle school. But for those students who had been suspended once during middle school, in fitted Model F, half of these students had been suspended a second time within 165 school days. Similarly, I estimate that half of the students who had been suspended twice were suspended a third time within 96 days of their return from a second suspension.

(d) Are the respective influences of gender, race, and poverty on the risk of suspension additive or interactive? Earlier in this paper, I cited national estimates of the suspension rate, cross-tabulated by gender and race. For instance, I described how 20% of Black boys, 12% of Black girls, 6% of White boys and 2% of White girls were suspended from U.S. public schools during the 2011-2012 school year (U.S. Department of Education Office for Civil Rights, 2014). Notice that, in the aggregate, these numbers counter the notion that Black *boys* are driving the suspension rate, when their suspension rate is only 67% higher than the suspension rate for Black girls. For White boys, nationally, the suspension rate—while only half the rate of the rate for Black girls—is *three times* the suspension rate for White girls.

These counter-intuitive effects were also manifested in my sample. Following the national trend for all public-school students, the cross-tabulated three-year suspension rate estimated for the Black and White middle-school students in my sample also demonstrates that Black *girls* were suspended at rates that far exceeded White students of either gender. During the three years of middle school, for instance, 63% of the Black

boys who started in district schools were suspended at least once, while the three-year suspension rate for Black girls was 46%. The corresponding three-year suspension rates for White students were 11.7% for boys and 3.3% for girls, respectively.

Nearly 80% of the Black students in my sample were eligible for free or reduced-price lunch, compared to less than 15% of White students, which raises the question of the extent to which poverty might account for racial differences in the risk of suspension. However, I found that poverty is only a partial explanation for the suspension rates in my sample. The three-year suspension rate for Black middle-school students who receive free/reduced-price lunch was 60%. But, for the relatively small number of Black students who do not report eligibility for this subsidy—fewer than 100 students from this sample cohort—the three-year suspension rate was 36%. Indeed, this three-year suspension rate was substantially higher than the 23% rate for White students who received free/reduced-price lunch. Only 5% of the White students who were not in poverty were suspended at all during the three years of middle school.

Ironically, then, the issue of statistical power doesn't arise due to the small number of non-poor Black students, but the small number of White students who were suspended repeatedly. The risk of suspension for non-poor White students—particularly for White females—confounded my efforts to adequately model the risk of subsequent suspensions adequately. *None* of the more than 350 White female students who were not eligible free/reduced-price lunch were suspended a third time. And so, I synthesized the findings for the first three models (Models D, E, and F) into a final model—Model G—that explored the joint effects of all three student characteristics of gender, race, and poverty, but retained the temporal features of risk distinguished in the answer to my first

research question (Model B). Thus, while remaining cognizant of the lack of statistical power to describe the risk for White females adequately, I fitted preliminary models that included the two-way and three-way interactions between these covariates, and interactions between these interactive effects and *SPELL*. Additionally, I was able to include a four-way interaction between *MALE*, *BLACK*, *FRPL* and *SPELL* for the first spell only, lacking statistical power to estimate the contribution of this term for the second and third spell.

I first fitted a "full" model in which I specified the risk of suspension by: *SPELL*; linear and quadratic *LNTIME*, and the two-way interaction of *LNTIME* with *SPELL*; *MALE*, and the two-way interaction of *MALE* with *SPELL*; *BLACK*, and the two-way interaction of *BLACK* with *SPELL*; *FRPL*, and the two-way interaction of *FRPL* with *SPELL*; the three-way interaction of *MALE*, *BLACK*, and *SPELL*; the three-way interaction of *MALE*, *FRPL*, and *SPELL*; the three-way interaction of *BLACK*, *FRPL*, and *SPELL*; and finally, the four-way interaction between *MALE*, *BLACK*, and *FRPL* for the first *SPELL* only. This model contained 29 parameters. After inspecting the results of fitting this "full" model, I trimmed the effect of *FRPL*, the interaction between *FRPL* and *MALE*, and the interaction between *FRPL* and *BLACK* for the second and third *SPELL*s (six parameters). I also found that the risk of a second or third suspension did not differ by *MALE* or the interaction between *MALE* and *BLACK* and their interaction, and so I trimmed four more parameters. Finally, I trimmed the three-way interaction between *MALE*, *BLACK*, and the first *SPELL*, and the four-way interaction between *MALE*, *BLACK*, *FRPL* and the first *SPELL*. Thus, in Model G, I fitted a "reduced" model with 12

fewer parameters, while yielding a deviance statistic of 14.6, for the saving of 12 degrees of freedom (p=0.26).

First, from the estimate of the slope parameter associated with the $MALE \times SP1$ interaction in Model G (1.50, p<.001), I find that the risk of a first suspension for a White non-poor male (compared to a White non-poor female) was substantially higher than the male/female difference in risk of first suspension that I found in Model D. Anti-logging the estimate for this slope parameter, I conclude that the fitted odds of a first suspension for White males not eligible for free/reduced-price lunch are 4.5 times the odds of a first suspension for corresponding White females. However, because interactions between gender and the predictors indicating second and third spell were not required in this model, I can conclude that there were no differences in the risk of suspension beyond the first spell for White, non-poor students.

Second, from the estimate of the slope parameter associated with the $BLACK \times SP1$ interaction in Model G (2.195, p<.001) I find that the risk of a first suspension for a non-poor Black female was still staggeringly high, compared to a non-poor White female. Anti-logging the estimate for this latter slope parameter, I conclude that the fitted odds of a first suspension for Black females not eligible for free/reduced-price lunch are *nine times* the odds of a first suspension for corresponding non-poor White females. Furthermore, once a non-poor female was suspended a first time, the risk of a second suspension also differed by race, as indicated by the estimated slope parameter associated with the $BLACK \times SP2$ interaction in Model G (0.601, p<.01). Anti-logging the estimate for this slope parameter, I conclude that the fitted odds of a second suspension for non-poor Black females are 1.82 times the odds for non-poor White females. I also conclude

that the fitted odds of a third suspension for non-poor Black females are 1.58 times the odds for non-poor White females, as indicated by the estimated slope parameter associated with the $BLACK \times SP3$ interaction in Model G (0.481, $p<.10$).

Third, from the estimated slope parameter associated with the $FRPL \times SP1$ interaction in Model G (2.423, $p<.001$), I estimate that the risk of a first suspension for a White female student who received free or reduced-price lunch (compared to a White non-poor female) was substantially higher than the difference in risk of first suspension estimated by the $FRPL \times SP1$ parameter that I found in Model E. Anti-logging the estimate for this slope parameter, I conclude that the fitted odds of a first suspension for White females who were eligible for free/reduced-price lunch are *11 times* the odds of a first suspension for corresponding non-poor White females. And again, because interactions between *FRPL* and the predictors indicating second and third spell were not required in this model, I can conclude that there were no differences in the risk of suspension for White females beyond the first spell.

Fourth, notice that I detected a three-way interaction between *MALE*, *FRPL*, and *SPELL* #1 in Model G. With this parameter—$MALE \times FRPL \times SP1$—I predict that the difference in the effect of *MALE* on the risk of a first suspension depends on students' eligibility for *FRPL*. Specifically, from the estimate of the slope parameter associated with the $MALE \times FRPL \times SP1$ interaction (-0.949, $p<.01$), I find that the gender difference in the risk of a first suspension for White students is moderated by whether a student is eligible for *FPRL*. By adding this (negative) effect of $MALE \times FRPL \times SP1$ to the "main effect" of $MALE \times SP1$ and anti-logging this combined estimate, I conclude that the gender difference in the fitted odds of a first suspension for White students is

"only" 1.74 for White students who were eligible for *FRPL*, compared to the above-described odds-ratio of 4.5 between males and females for non-poor White students.

Fifth, also notice that I detected a three-way interaction between *BLACK*, *FRPL*, and *SPELL* #1 in Model G. With this parameter—$BLACK \times FRPL \times SP1$—I predict that the difference in the effect of *BLACK* on the risk of a first suspension depends on students' eligibility for *FRPL*. Specifically, from the estimate of the slope parameter associated with the $BLACK \times FRPL \times SP1$ interaction (-0.929, $p<.01$), I find that the racial difference in the risk of a first suspension for females is moderated by whether a student is eligible for *FPRL*. By adding this (negative) effect of $BLACK \times FRPL \times SP1$ to the "main effect" of $BLACK \times SP1$ and anti-logging this combined estimate, I conclude that the racial difference in the fitted odds of a first suspension for female students is "only" 3.5 for Black females who are eligible for *FRPL*, compared to the above-described odds-ratio of 9.0 between non-poor Black females and non-poor White females.

To better tease out and display these complex joint effects, in Figure 7, I present a plot of the fitted hazard (top panel) and the survival (bottom panel) functions for Model G, highlighting these results for prototypical students at "average" levels of free or reduced-price lunch status.[24] Again, notice that the risk of suspension in Spell #1 was lower throughout middle school, compared to the risk of a second or third suspension, if such events occurred. I estimate that the risk of a first suspension peaked on school day 66, and then slowly diminished throughout the rest of middle school. The profile of risk

[24] The overall proportion of students eligible for free or reduced-price lunch in my sample was 43%.

for a second suspension was substantially higher than a first suspension. It was notably higher immediately following a first suspension, and the risk was 82% higher for Black students, compared to White students. The estimated risk of a third suspension was highest on the fifth day following students' return to school following their previous (second) suspension. Furthermore, I estimate that the risk of a third suspension for Black students was 58% higher than the corresponding risk for a White student.

In the bottom panel, I provide a corresponding fitted survival plot from the start of the first three spells, for Black students and White students of each gender, at "average" levels of *FRPL*. This plot is the compilation of eight different estimates of risk. First, note that the risk of first suspension differed by race, gender and their interaction. At "average" levels of poverty, 95% of prototypical White females "survived" middle school without being suspended, as denoted by the solid black line. For prototypical White males, I estimate that the controlled survival rate—across 529 days of middle school—was 86%. In contrast, the survival rate for prototypical Black female students was 63%, and survival rate for Black male students was 25%.

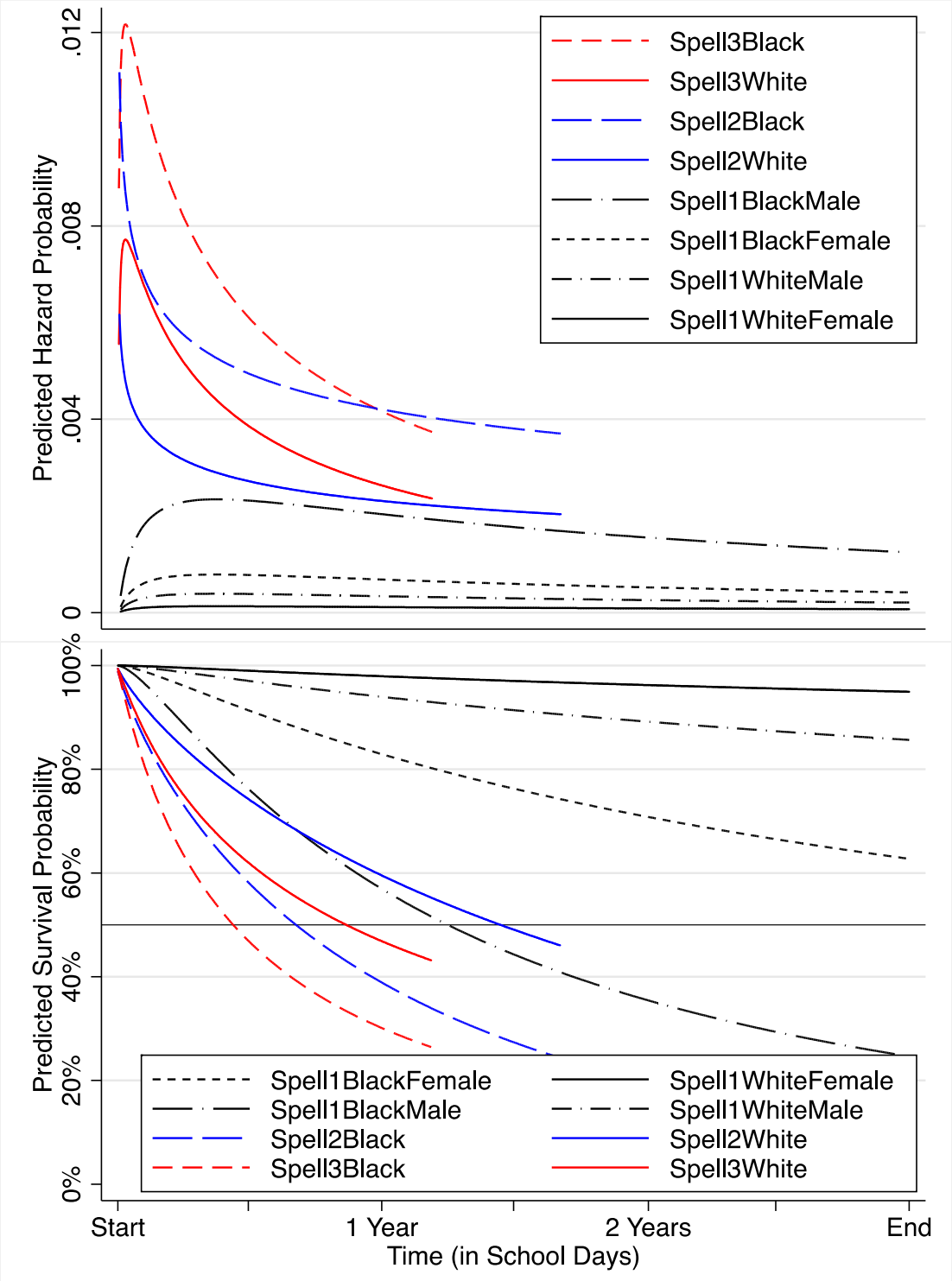

Figure 7. Predicted hazard probability (top panel) and survival probability (bottom panel) of a students' first, second and third suspension from middle school, Black students and White students, at average levels of free or reduced-price lunch status ($n_{students}$=1,274, $n_{suspensions}$=613). Predicted values obtained from fitting Model G in Table 3.

Regarding the middle-school students who had been suspended a first time, again note that the fitted survival probability until a subsequent suspension also differed by race. I estimate that the median "lifetime" of a prototypical White student who had been suspended a first time in middle school, controlling for *FRPL*, was approximately 257 school days until a second suspension. The estimated median "lifetime" for a prototypical Black student who had been suspended a first time was approximately 120 school days until a second suspension from middle school. A White student who had been suspended a second time "survived" approximately 154 school days until a third suspension from middle school occurred, while my corresponding predicted survival probability for Black students was only 78 days.

Threats to Validity

My study highlights two main stories: that the risk of subsequent suspensions is substantially higher than the risk of a first suspension; and that the risk of each suspension—both the first and subsequent suspensions—is higher for Black students than for White students, even after controlling for students' level of poverty, as measured by their eligibility for free or reduced-price lunch. However, I describe patterns of out-of-school suspensions in only one school district, in one region of the country. I cannot generalize my findings to public schools in other locations, as there are likely demographic and district characteristics of this case study that influence my findings directly. Thus, I am cautious about making broad claims regarding about my research.

My study is also situated in one three-year time period, from September 2007 through June 2010. In retrospect, during this period, the Great Recession was at its lowest economic point. This time period is also bisected by the election of President Obama, and

it predates the turmoil of an increasingly divided government. Indeed, contentious debates about collective bargaining for public-school teachers began in earnest after 2010 in many states, and they continue at the time of this writing. Additionally, the *Black Lives Matter* movement has highlighted troubling racial disparities in law enforcement that may also have parallels in how school discipline is administered in schools. Thus, patterns in school suspension may be different today than they were six or eight years ago.

My dataset also did not include any information about students' suspension record in elementary school. For instance, I did not know—and could not incorporate into my analyses—whether a student recorded as being "first suspended" in 6th grade had actually been suspended earlier, perhaps in 4th grade. Consequently, my results may be different, had I been able to incorporate the students' elementary-school discipline history into account. But to reconstruct a profile that describes the risk of suspension across an entire school career would require nine years of longitudinal data, as I would need to follow each student from their entry in kindergarten through the end of middle school. I hope to adopt this approach in my future research, with new longitudinal data.

Another threat to the validity of the current analysis was that the students in this sample attended middle schools with different school-wide suspension rates. As I noted earlier, I included a random intercept to represent differences in the risk of suspension associated with attending a particular district middle school. These random effects make a statistically significant contribution in all of my fitted models, with intra-class correlations of between 2.2% and 4.5% for the various "final" fitted models. Because these intra-class correlations at that range are rather small, I refitted my models without

the corresponding random effects and found nearly identical parameter estimates for each corresponding model (using "robust" standard errors). The results of fitting these additional models can be found in Table A5, in the Appendix. I also refitted Models B, D, E, F, and G with an additional fixed parameter included to distinguish the specific risks of attending each school. Again, as I show in Table A6 (in the Appendix), the obtained parameter estimates are very similar.

But does this imply that perhaps one school with very high suspension rates drove my results? In order to test the sensitivity of my findings to this conjecture, I compared the results of Models B, D, E, and F with the results of fitting the same model, but after dropping the school with the highest school-wide suspension rate. In this sensitivity check, I found that my results were slightly different in magnitude, but the direction and almost all of the statistical significance of the findings remained unchanged.

For example, in Figure A1 (in the Appendix), I present a plot of fitted hazard function (top panel) and fitted survival function (bottom panel) for Model 23T3—the most complex of the models in my analysis—as part of my sensitivity analysis. While the school that I removed from this sensitivity analysis had a slightly lower level of poverty (41%), and a higher suspension rate, the underlying survival probabilities associated with gender, race, and poverty remain. The median survival time to a second suspension for a prototypical white student from the reduced sample was substantially shorter—199 days, compared to 257 days for the full model. And the time to a second suspension for a prototypical Black student was a bit longer (although still substantially shorter than the estimate for White students)—132 days in this reduced model, compared to 120 days in the full model. The time to a third suspension for prototypical White students was 134

days in the reduced model, compared to 154 days in the full model. My estimates of the median time until a third suspension for prototypical Black students were 84 days in the reduced model, compared to 78 days in the full model.

Regarding the smaller sample that I used to estimate the risk of suspension in models E and G, this compromise might also have affected my findings. As I explained, in the preceding **Measures** section, I was forced to remove Asian, Hispanic, and Native American students from the samples for which I fitted any models that included race/ethnicity as a predictor, as too few students in those categories populated the risk set for subsequent spells. In my previous work, with larger sample sizes and less complexity (Hoffman, 2014a; Hoffman, 2014b; Hoffman, 2014c), I found that Hispanic students were at greater risk of school discipline (both out-of-school suspensions and expulsions), but that the risk for Asian students was lower, compared to White students. However, the demographic composition of public schools is changing rapidly, with many schools enrolling increasing numbers of Asian and Hispanic students, perhaps making finer-grained estimates of the risk of suspension plausible in future studies.

Many states (including the state in which my study is situated) have broadened the types of racial/ethnic classification of students, making statistical analyses perhaps more difficult. The five self-reported racial/ethnic categories that the state requested of parents during the period of my study (data typically collected by school secretaries) are now supplemented by a category allowing families to indicate that students identify as bi- or multi-racial, as well as splitting the category “Asian” into “Asian and Native Hawaiian” or “Other Pacific Islander”. Furthermore, Hispanic students (a term rather out of favor now, though still used by the U.S. census) can be of any race—as they always

have been. But most studies still classified Hispanic students in a category distinct from Black students, even though people of Haitian origin and Cuban origin are classified together. Grouping students into three non-overlapping categories—White, non-Hispanic; Hispanic; and Black—may be increasingly untenable. And simple comparisons of risk between Black students and non-Hispanic White students will be likely inadequate for my future research.

Finally, by design, I did not take into account the date of a first suspension—and subsequently, the date marking the beginning of a second spell—in the risk of a second suspension. In other words, my estimates of the risk of a second suspension do not include any information about *when* the first day of the second spell occurred. Nor had the risk of a third suspension include information about when the beginning of the third spell occurred. In preliminary work, I considered models that included a coarse estimate of time, with a fixed reference point estimating the total time that a student spent from the beginning of middle school until a suspension event, in any numbered spell. In this alternative strategy, I chose to estimate the risk of each of the numbered suspensions *yearly* (if a suspension occurred). However, because there is insufficient statistical power in my dataset to include both *year* and polynomial specifications of time, I chose to report the results of only one strategy in the main body of the text.

In Table A8, however, I present the results of this alternative specification of the risk of a first, second, or third suspension, by year for the five final models (models B, D, E, F, & G). Notice, first, that the parameter estimates for describing the baseline risk of suspension in spells 1, 2, and 3 in each of the five "final" models are substantially different than those in my preferred models in tables 2 and 3. However, also note that the

parameter estimates for the critical covariates in each of the models associated with models D, E, F, & G are very similar. The only aspect of this alternative strategy that does not produce similar risk functions is for those that describe differences in the risk of a second or third suspension between Black students and White students. I attribute these differences to the fact that many students—particularly Black students—were suspended a first, second, and third time during 6th grade. Unfortunately, under this alternative strategy, I must assume that the risk function is constant for the entirety of each grade. In my preferred modeling strategy, presented in the main body of the text, I demonstrate that the risk functions differ substantially, over time.

Discussion

As I noted in my introductory comments, the work of school principals includes dealing daily with the impulsive and erratic behaviors of students. More than 40 years ago, the U.S. Supreme Court noted that public-school principals had the authority—perhaps even the obligation—to suspend students out-of-school for acts of misbehavior (*Goss v. Lopez*, 1975). Principals' vigilant attention to order and discipline requires them to weigh both the facts of the situation and the policies and procedures that serve as guides to suspending students. I believe, though, that policymakers, school superintendents, and school boards have not given the same vigilant attention to the task. Simply put, school personnel are responsible for dealing with student behavior, and documenting staff actions to the district office and the state. Yet these reports are typically overlooked. In this study, I explain how school suspensions might be assessed in a productive and meaningful way.

The Suspension Rate

I first highlight similarities and differences between my results and the traditional suspension rate. In this study, 11% of the sampled students experienced a first suspension in middle school during 6th grade, while attending district schools. This simple statistic—187 first suspensions, for the 1,693 sample students—is roughly comparable the 12.1% "suspension rate" reported to the public by the school district. However, there are subtle differences. The rate that I estimate here is a simple count of the number of first suspensions for my sampled students occurring in 6th grade in district middle schools. In contrast, the state's methodology—reported publicly—counts first suspensions for all 6th grade students suspended, including students who move in or out of the school (mobile students) in the numerator, but uses as its denominator the enrollment count on the third Friday of September. (This student count is used in calculating state funding.) And so, notice that by the end of 6th grade, my sample is smaller. More than 50 of the original students who started middle school in district schools had moved away during 6th grade, and two were expelled.

However, using survival-analytic techniques, I am then able to describe the survival rate until a first suspension from middle school in 7th grade for sampled students who had not been suspended in 6th grade. I estimate that about 82.5% survived both 6th and 7th grade, and, correspondingly, another 6.5% of sample students who had not been suspended in 6th grade were suspended in 7th grade. Similarly, 56 students survived 6th and 7th grade, only to be suspended in 8th grade for their first time in middle school. As I explained in the **Repeated Suspensions from School** section above, publicly reported

suspension rates cannot provide any information about whether students who were suspended in 6th grade were suspended again in 7th grade. Using survival analysis can.

Employing survival-analytic techniques over multiple years, I can estimate a multi-year "suspension rate" that differs from the rate for each single year. Keep in mind that only 1,261 students (74% of the original sample) were still at risk of a first suspension at the beginning of 8th grade, as the others had either been suspended during 6th or 7th grade or had moved away. The state-reported summaries of whether students were suspended, by grade level—12.1% for 6th grade, 14.0% for 7th grade, and 14.6% for 8th grade—can now be also described as a single statistic: 20% of the cohort of students who began 6th grade in the district were suspended at least once before they either completed 8th grade three years later or moved to a different school district. My results, then, can be compared to the techniques used by Petras, et al. (2011), when they found that 22.8% of Baltimore elementary students were suspended at least once during 1st through 7th grade, and Fabelo et al. (2011), who found that 31% of all students were suspended at least once during secondary school.

Furthermore, by estimating the risk of a first suspension within this framework—with discrete time periods measured in school days—I am also able to provide sensible estimates of when this risk is higher or lower, during students' middle-school career. I show, explicitly, in Figure 3, that the risk of a first suspension was highest during the middle of 6th grade (perhaps long enough for students to acclimate to secondary-school). But this risk tapers off only gradually. Thus, in this case study, I provide additional evidence that the first out-of-school suspension for many students occurred somewhat randomly throughout their entire middle-school career, as I first noted in my previous

work employing Cox-regression analysis (Hoffman, 2014a). By the end of 8th grade, the risk of a first suspension had diminished by more than two-thirds. Some students had been suspended and were now assigned to a second spell. Still, students were still at risk of a first suspension (and, indeed, were suspended) even after they have spent 500 school days in middle school without being suspended previously.

Thus, policymakers might consider unintended consequences from lengthening the school year. For example, if we assume that patterns of risk would remain comparable for days beyond the roughly 177 days of the school year, I postulate that a longer school year would correspond with an increase in the overall suspension rate, purely as a function of the increased time at risk. Each day that a student is eligible to attend middle school is essentially another day that a student is at risk of suspension. So, in this sample, I estimate that an additional student would have been suspended (who had not been suspended before) on every additional third day of attendance.

Multiple Spells

Losen and Skiba (2010) noted that acts of misbehavior in middle school are, in many ways, typical adolescent behaviors. Rule breaking that leads to an out-of-school suspension is common. Yet, standard record keeping may not fully capture the extent of the issue. Many students are suspended repeatedly, making the suspension rate a conservative estimate of the frequency of suspensions, particularly suspensions experienced within a particular demographic group. Furthermore, as Lortie (2009) pointed out, some behavioral norms are established early. Thus, while my estimate of the high point in the risk profile of first suspension during the fall of 6th grade provided a

glimpse of a largely unstudied phenomenon, coarse estimates of the suspension rate disguise important subtlety.

Yes, some students were suspended exactly one time during middle school. In this study, I found that more than one-third of these students managed to keep their school record somewhat clean, with just one suspension in district middle schools. While the risk of subsequent suspensions was quite high, perhaps a single suspension could be explained as a misunderstanding. Other times, an incident was followed by their parents' forceful reaction, making it clear that no further incidents would be tolerated.

Still, even my estimate that more than one-third of the students were not suspended again is a conservative estimate. As I explained earlier, about 14% of sample students left the district before completing 8th grade. It is likely that some of them were suspended a second time in middle school, for an incident that occurred in another district. Furthermore, I had decided, in my analysis, that students who had been expelled from the school district were excluded from further analysis. At least some of the seven who were expelled in connection with their first suspension incident were likely to be suspended again during middle school, if they had returned after being excluded for a longer period of time—perhaps until the end of the school term.

However, my analysis illuminates substantial differences in risk between first and subsequent suspensions. More than 60% of the sampled students who had been suspended a first time were suspended again in district middle schools. And by employing a multiple-spell, discrete-time survival analysis framework, I demonstrate that—for a student who was suspended a first time—the risk of a second suspension was quite high. As I illustrated in the predicted survival probability plot, in Figure 3, "You

again?" might be a typical response by a principal. As I have shown, students were suspended for a second disciplinary incident, often within a matter of days after a student had returned following a first suspension.

Thus, my results from fitting Model B, to address my first research question, were somewhat disheartening to me, as a former middle school principal. I certainly knew that the discipline incidents that I faced on a typical school day included episodes of misbehavior by students that I had suspended before. But perhaps I had the romantic notion that the combination of swift punishment, enlisting the help of their parents, and a stern talking-to during the re-admit conference upon return would dissuade students from doing something to get themselves suspended again. My analysis shows otherwise. Students across the district displayed the same behavioral pattern, with most students suspended again. Furthermore, the time until a third suspension was even shorter, on average, than it was for a second suspension.

Student Gender

If a principal was sitting at her desk and a student arrived at the office to be disciplined for a first substantial offense, the principal would probably guess that the student coming through the door is more likely to be male than female. In this study, I found that the odds that the student entering the office was male were 84% higher than the odds that the student was female. My results, then, correspond with scholarly research going back to at least the 1970s. As Bertrand and Pan (2013) memorably asserted in "The trouble with boys", disruptive behavior in schools likely has an inherent biological aspect.

Suppose, however, that the principal had been informed that someone was here to be disciplined and the student had already suspended once. Was the pool of candidates for a second suspension also predictable by student gender? In my analysis, I found that the conditional probability of a second suspension did not depend on the gender of the student. Nor did the conditional probability of a third suspension depend on gender. Rather, I show that there was a set of students—about twice as many boys as girls—who had been suspended. And once they'd been suspended, the risk of a second suspension wasn't predictable based on whether the student was female or male. Yet, while this phenomenon holds true for gender (and also for differences by free or reduced-price lunch status), it does not hold true for race.

Student Race

Why are Black students, as a demographic group, suspended at rates an order of magnitude higher, compared to White students. Most researchers concluded that differences in the rate of suspension, by race, were attributed to pervasive differences in the school experiences of Black and White students. For example, Gregory, Cornell, and Fan (2011) combined their analysis of Virginia state records of school suspensions with the results from a survey of 9th grade students, sampled from more than 90% of the high schools in Virginia. They concluded that suspension rates for Black students and White students, which differ substantially, were not attributable to differences in either school size or SES. Rather, their findings are more consistent with those of Ferguson (2000) and Payne and Welsh (2010)—that the school experience for Black students is decidedly different, with Black students being subject to more punitive discipline than other students who committed the same offenses.

What my study adds to this literature is that Black students—as a group—were at higher risk for subsequent suspensions, beyond the already high risk of a first suspension. That is to say, the independence regarding the risk of second and third suspensions that I found for gender did not hold true for this estimate of conditional risk between Black students and White students in my analysis. For students who had been suspended once, the risk of a second suspension also differed by race; the odds of a second suspension for a Black were 82% higher than the odds for a White student. This phenomenon continued until a 3rd suspension, where the risk for Black students was 58% higher than the risk for White students.

School personnel in this study were likely aware already that Black students were more frequently in the office for discipline. But what I find truly disconcerting is that the time between suspensions was different, based on race. A White student who had been suspended might "survive" more than 250 days after returning from a first suspension before getting trouble again. But a Black student returning from a first suspension is predicted to survive only half as long!

The Joint Effects of Gender, Race, and Poverty

Few studies have explored whether differences in school discipline by gender also depend on race. However, as I noted earlier, national statistics on school suspension clearly indicate a gender-race interaction. Nationally, Black boys were suspended at a rate two-thirds higher than Black girls, while White boys were suspended at a rate three times higher, during the 2011-12 school year (U.S. Department of Education Office for Civil Rights, 2014). And, certainly the results of my analysis also point to a gender-by-race interaction. The risk of first suspension for Black male students were the highest of

any racial/ethnic and gender demographic group—three times the risk of White males. However, my analysis also corroborates, to an extent, the findings of Morris (2005), who observed, in a case study of one Texas school, that Black female students were likely to be disciplined at rates similar to Black male students.

In Model G, I found that the risk of a first suspension was substantially lower for White students who were not eligible for free or reduced-priced lunch. For White, non-poor females, this risk was very small, estimated by only a handful of suspensions in this analysis. The risk for White, non-poor males, was 4.5 times larger—accounting for most of the suspensions for non-poor Whites. But for the few of these non-poor White students who were suspended a first time, the risk was substantially higher, but didn't differ by gender.

For White students, I found an important statistical interaction between gender and poverty. The difference in the odds of a first suspension for White female students in poverty was *eleven times* the risk for non-poor White female students. This difference was among the highest disparities among demographic groups. Essentially, almost no non-poor White girls were suspended. But once White females of any social class had been suspended once, they were at substantial risk of a subsequent suspension.

In contrast, Black students were at substantially higher risk than White students—for both males and females—regardless of their eligibility for free/reduced-priced lunch. This phenomenon is particularly striking for non-poor Black females, as their risk of a first suspension is *more than nine times the risk* for a non-poor White female. Furthermore, the odds of a second suspension for non-poor Black females were 82%

higher than the odds of a second suspension for non-poor White females—a finding that isn't documented on other research.

However the "full" model, presented as Model G and Figure 7, highlights what I believe was my most provocative insight: The risk of suspension for Black students was substantially higher than for White students, even after controlling for poverty and suspension history. Skiba et al. (2011) explained that despite the widespread belief that much of the discipline gap can be attributed to poverty, the overrepresentation of Black students in school disciplinary outcomes cannot be fully explained by any other variable than race itself. My analysis confirms this, and adds to a large body of studies about disciplinary outcomes between Black and White students. This racial disparity has been on the agenda of researchers since Edelman (1975) and her colleagues brought this issue to the fore 40 years ago.

References

Arum, R. (2003). *Judging school discipline: The crisis of moral authority*. Cambridge, MA: Harvard University Press.

Arum, R., & Preiss, D. (2009). Law and disorder in the classroom. *Education Next, 9*(4)

Aud, S., Wilkinson-Flicker, S., Kristapovich, P., Rathbun, A., Wang, X., & Zhang, J. (2013). *The condition of education.* (No. NCES 2013-037). Washington, D.C.: U.S. Department of Education, National Center for Education Statistics.

Autor, D. H., Figlio, D. N., Karbownik, K., Roth, J., & Wasserman, M. (2015). *Family disadvantage and the gender gap in behavioral and educational outcomes*. Northwestern University: Institute for Policy Research.

Beck, A. N., & Muschkin, C. G. (2012). The enduring impact of race: Understanding disparities in student disciplinary infractions and achievement. *Sociological Perspectives, 55*(4), 637-662. doi:10.1525/sop.2012.55.4.637

Bertrand, M., & Pan, J. (2013). The trouble with boys: Social influences and the gender gap in disruptive behavior. *American Economic Journal: Applied Economics, 5*(1), 32-64.

Cook, R. J., & Lawless, J. F. (2007). *The statistical analysis of recurrent events*. New York, NY: Springer.

Cox, D. R. (1972). Regression models and life-tables (with discussion). *Journal of the American Statistical Association, Series B, 34*, 187-220.

Department of Public Instruction. (2013). *Wisconsin's information network for successful schools.*

Edelman, M. W., Beck, R., & Smith, P. V. (1975). *School suspensions--are they helping children?* Cambridge, Mass.: Children's Defense Fund.

Fabelo, T., Thompson, M. D., Plotkin, M., Carmichael, D., Marchbanks, M. P. I., & Boothe, E. A. (2011). *Breaking schools' rules: A statewide study of how school discipline relates to students' success and juvenile justice involvement.* New York, NY; College Station, TX: Council of State Governments Justice Center; Public Policy Research Institute of Texas A&M University.

Ferguson, A. A. (2000). *Bad boys: Public schools in the making of black masculinity*. Ann Arbor, MI: The University of Michigan Press.

Goss v. Lopez, 419 U.S. 565, Justia US Supreme Court Center (U.S. Supreme Court 1975).

Gregory, A., Cornell, D., & Fan, X. (2011). The relationship of school structure and support to suspension rates for black and white high school students. *American Educational Research Journal, 48*(4), 904-934.

Hoffman, S. L. (2014a). *Estimating the risk of school suspension: A case study in one county's public middle schools.* (Presentation). Trumbull, CT: 45th Annual Meeting of the Northeastern Educational Research Association.

Hoffman, S. L. (2014b). *Improving the estimation of the risk of school suspension using continuous-time survival analysis: A case study in the public middle schools in one metropolitan region.* (Unpublished Qualifying Paper). Harvard Graduate School of Education, Cambridge, MA.

Hoffman, S. L. (2014c). Zero benefit: Estimating the effect of zero tolerance discipline polices on racial disparities in school discipline. *Educational Policy, 28*(1), 69-95. doi:10.1177/0895904812453999

Kaplan, E. L., & Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association, 53*(282), 457-481.

Kinsler, J. (2011). Understanding the black-white school discipline gap. *Economics of Education Review, 30*(6), 1370-1383. doi:10.1016/j.econedurev.2011.07.004

Lortie, D. C. (2009). *School principal: Managing in public*. Chicago: University of Chicago Press.

Losen, D., & Skiba, R. J. (2010). *Suspended education: Urban middle schools in crisis.* Southern Poverty Law Center.

Masyn, K. E. (2009). Discrete-time survival factor mixture analysis for low-frequency recurrent event histories. *Research in Human Development, 6*(2-3), 165-194. doi:10.1080/15427600902911270

McCarthy, J. D., & Hoge, D. R. (1987). The social construction of school punishment: Racial disadvantage out of universalistic process. *Social Forces, 65*(4), 1101-1101-1120.

McFadden, A. C., & Marsh II, G. E. (1992). A study of race and gender bias in the punishment of school children. *Education & Treatment of Children (ETC), 15*(2), 140.

Morris, E. W. (2005). “Tuck in that shirt!” race, class, gender, and discipline in an urban school. *Sociological Perspectives, 48*(1), 25-48.

Ornstein, A. (1982). Student disruptions and student rights: An overview. *The Urban Review, 14*(2), 83-91.

Payne, A. A., & Welch, K. (2010). Modeling the effects of racial threat on punitive and restorative school discipline practices. *Criminology, 48*(4), 1019-1062. doi:http://dx.doi.org.ezp-prod1.hul.harvard.edu/10.1111/j.1745-9125.2010.00211.x

Petras, H., Masyn, K. E., Buckley, J. A., Ialongo, N. S., & Kellam, S. (2011). Who is most at risk for school removal? A multilevel discrete-time survival analysis of individual- and context-level influences. *Journal of Educational Psychology, 103*(1), 223-237. doi:10.1037/a0021545

Rabe-Hesketh, S., & Skrondal, A. (2012). *Multilevel and longitudinal modeling using Stata* (3rd ed.). College Station, Tex.: Stata Press Publication.

Raffaele Mendez, L. M., & Knoff, H. M. (2003). Who gets suspended from school and why: A demographic analysis of schools and disciplinary infractions in a large school district. *Education and Treatment of Children, 26*(1), 30-30-51.

Singer, J. D., & Willett, J. B. (1993). It's about time: Using discrete-time survival analysis to study duration and the timing of events. *Journal of Educational Statistics, 18*(2), 155-195.

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. Oxford; New York: Oxford University Press.

Skiba, R. J., Horner, R. H., Choong-Geun Chung, Rausch, M. K., May, S. L., & Tobin, T. (2011). Race is not neutral: A national investigation of African American and Latino disproportionality in school discipline. *School Psychology Review, 40*(1), 85-107.

Skiba, R. J., Peterson, R. L., & Williams, T. (1997). Office referrals and suspension: Disciplinary intervention.. *Education & Treatment of Children (ETC), 20*(3), 295.

Sullivan, A. L., Klingbeil, D. A., & Van Norman, E. R. (2013). Beyond behavior: Multilevel analysis of the influence of sociodemographics and school characteristics on students' risk of suspension. *School Psychology Review, 42*(1), 99-114.

Theriot, M. T., & Dupper, D. R. (2010). Student discipline problems and the transition from elementary to middle school. *Education and Urban Society, 42*(2), 205-222. doi:10.1177/0013124509349583

Townsend, B. L. (2000). The disproportionate discipline of African American learners: Reducing school suspensions and expulsions. *Exceptional Children, 66*(3), 381-381-391.

U.S. Department of Education. (2012). *Digest of educational statistics*. Washington DC: U.S. Dept. of Health, Education, and Welfare, Office of Education.

U.S. Department of Education. (2014). *Guiding principles: A resource guide for improving school climate and discipline.* (). Washington, D.C.:

U.S. Department of Education Office for Civil Rights. (2014). *Civil rights data collection: Data snapshot (school discipline).* (No. 1). Washington D.C.:

Wallace, J.,John M., Goodkind, S., Wallace, C. M., & Bachman, J. G. (2008). Racial, ethnic, and gender differences in school discipline among U.S. high school students: 1991-2005. *Negro Educational Review, 59*(1), 47-62.

Willett, J. B., & Singer, J. D. (1995). It's déjà vu all over again: Using multiple-spell discrete-time survival analysis. *Journal of Educational and Behavioral Statistics, 20*(1), 41-67.

Wu, S., Pink, W., Crain, R., & Moles, O. (1982). Student suspension: A critical reappraisal. *The Urban Review, 14*(4), 245-303. doi:10.1007/BF02171974

Appendix—Supplementary Tables and Figures

Table A1. Parameter estimates, approximate *p*-values, and goodness-of-fit statistics from five fitted multiple-spell hazard models predicting the risk of out-of-school suspension by spell, time within spell, and student gender ($n_{students}$=1,693, $n_{suspensions}$=691).

	Model 4	Model 5	Model 6	Model 7	Model 8
SPELL1	-16.837	-10.544***	-11.086***	-10.682***	-10.682***
SPELL2	-4.796***	-5.232***	-5.304***	-5.044***	-4.864***
SPELL3	-5.538***	-4.630***	-5.163***	-4.994***	-4.929***
LNTIME×SP1	4.413	0.992	1.615***	1.590***	1.590***
LNTIME×SP2	-0.458	-0.077	-0.051	-0.062	-0.061
LNTIME×SP3	0.821	-0.391	0.363	0.381	0.377
LNTIME2×SP1	-0.6	0	-0.194***	-0.204***	-0.204***
LNTIME2×SP2	0.085	-0.02	-0.01	-0.021	-0.022
LNTIME2×SP3	-0.2	0.179	-0.085	-0.102*	-0.101*
LNTIME3×SP1	0.016	-0.018			
LNTIME3×SP2	-0.007	0.002			
LNTIME3×SP3	0.006	-0.027			
MALE×SP1	7.115	0.542	1.104	0.613***	0.612***
MALE×SP2	-0.685	0.059	0.608	0.274	
MALE×SP3	1.064	-0.211	0.348	0.096	
MALE×LNTIME	-2.575	0.752	0.023		
MALE×LNTIME2	0.263	-0.264	-0.024		
MALE×LNTIME3	-0.003	0.023			
MALE×LNTIME×SP2	4.003				
MALE×LNTIME×SP3	1.605				
MALE×LNTIME2×SP2	-0.719				
MALE×LNTIME2×SP3	0.013				
MALE×LNTIME3×SP2	0.043				
MALE×LNTIME3×SP3	-0.021				
rho	0.047	0.046	0.046	0.047	0.047
rank	25	19	15	13	11
-2LL	10029.2	10036.0	10038.0	10043.2	10047.2
LR Test (compared to Model 4)		6.782	8.855	13.996	17.899
df		6	10	12	14
p-value		0.34	0.55	0.30	0.21

*~ p<0.10, * p<0.05, ** p<0.01, *** p<0.001*

Table A2. Parameter estimates, approximate *p*-values, and goodness-of-fit statistics from six fitted multiple-spell hazard models predicting the risk of out-of-school suspension by spell, time within spell, and student gender ($n_{students}$=1,274, $n_{suspensions}$=613).

	Model 9	Model 10	Model 11	Model 12	Model 13	Model 14
SPELL1	-15.59**	-10.50***	-9.98***	-11.08***	-11.08***	-11.08***
SPELL2	-4.46***	-5.18***	-4.10***	-5.07***	-5.07***	-4.60***
SPELL3	-4.23***	-4.79***	-4.05***	-5.08***	-4.71***	-4.70***
LNTIME×SP1	5.81	1.71	0.88	1.29**	1.29**	1.29**
LNTIME×SP2	-0.67	1.14	-0.56	-0.14	-0.14	-0.12
LNTIME×SP3	0.2	0.94	-0.18	0.28	0.32	0.32
LNTIME2×SP1	-1.46	-0.41	-0.12*	-0.16***	-0.16***	-0.16***
LNTIME2×SP2	0.26	-0.58	0.03	-0.01	-0.01	-0.01
LNTIME2×SP3	-0.15	-0.43	-0.04	-0.09	-0.10*	-0.09*
LNTIME3×SP1	0.11	0.03				
LNTIME3×SP2	-0.04	0.06*				
LNTIME3×SP3	0.01	0.04				
BLACK×SP1	7.77	1.75	0.93	2.33***	2.33***	2.32***
BLACK×SP2	-0.74	0.19	-0.64	0.60**	0.59**	
BLACK×SP3	-0.62	0.1	-0.73	0.46~		
BLACK×LNTIME	-5.62	-0.69	0.52			
BLACK×LNTIME2	1.64	0.37	-0.05			
BLACK×LNTIME3	-0.14	-0.04				
BLACK×LNTIME×SP2	7.14					
BLACK×LNTIME×SP3	5.86					
BLACK×LNTIME2×SP2	-2.28					
BLACK×LNTIME2×SP3	-1.63					
BLACK×LNTIME3×SP2	0.22*					
BLACK×LNTIME3×SP3	0.14					
rho	0.02	0.02	0.02	0.02	0.02	0.02
rank	25	19	15	13	12	11
-2LL	8393.2	8402.8	8407.1	8411.9	8415.3	8425
LR Test (compared to Model 9)		9.578	13.868	18.737	22.102	31.817
df		6	10	12	13	14
p-value		0.144	0.179	0.095	0.054	0.004

*~ p<0.10, * p<0.05, ** p<0.01, *** p<0.001*

Table A3. Parameter estimates, approximate *p*-values, and goodness-of-fit statistics from five fitted multiple-spell hazard models predicting the risk of out-of-school suspension by spell, time within spell, and free/reduced-price lunch status ($n_{students}$=1,693, $n_{suspensions}$=691).

	Model 15	Model 16	Model 17	Model 18	Model 19
SPELL1	-9.32***	-11.20***	-10.37***	-11.27***	-11.27***
SPELL2	-7.82***	-5.41***	-4.29***	-5.08***	-4.84***
SPELL3	-3.85***	-4.88***	-4.22***	-5.02***	-4.91***
LNTIME×SP1	0.4	2.22	1.12*	1.48***	1.49***
LNTIME×SP2	3.39	1.22	-0.42	-0.07	-0.06
LNTIME×SP3	-0.36	0.89	0	0.37	0.38
LNTIME2×SP1	-0.03	-0.51	-0.15**	-0.19***	-0.19***
LNTIME2×SP2	-1.09	-0.56	0.01	-0.02	-0.02
LNTIME2×SP3	0.08	-0.35	-0.07	-0.10*	-0.10*
LNTIME3×SP1	0	0.03			
LNTIME3×SP2	0.09	0.06			
LNTIME3×SP3	-0.02	0.03			
FRPL×SP1	-1.29	1.55	0.55	1.71***	1.71***
FRPL×SP2	2.97	0.25	-0.74	0.31	
FRPL×SP3	-1.28	0.14	-0.87	0.14	
FRPL×LNTIME	1.74	-0.91	0.47		
FRPL×LNTIME2	-0.27	0.42	-0.05		
FRPL×LNTIME3	0.01	-0.04			
FRPL×LNTIME×SP2	-5.1				
FRPL×LNTIME×SP3	-0.97				
FRPL×LNTIME2×SP2	1.29				
FRPL×LNTIME2×SP3	0.12				
FRPL×LNTIME3×SP2	-0.1				
FRPL×LNTIME3×SP3	0				
rho	0.033	0.033	0.033	0.033	0.033
rank	25	19	15	13	11
-2LL	9853.8	9860.6	9865	9868.6	9871.9
LR Test (compared to Model 15)		6.735	11.154	14.723	18.011
df		6	10	12	14
p-value		0.346	0.346	0.257	0.206

*~ p<0.10, * p<0.05, ** p<0.01, *** p<0.001*

Table A4. Parameter estimates, approximate *p*-values, and goodness-of-fit statistics from five fitted multiple-spell hazard models predicting the risk of out-of-school suspension by spell, time within spell, gender, race, and free/reduced-price lunch status ($n_{students}$=1,274, $n_{suspensions}$=613).

	Model 20	Model 21	Model 22	Model 23
SPELL1	-13.07***	-13.07***	-13.07***	-12.55***
SPELL2	-5.899***	-5.133***	-5.083***	-5.084***
SPELL3	-5.821***	-5.437***	-5.078***	-5.079***
LNTIME×SP1	1.226**	1.227**	1.227**	1.225**
LNTIME×SP2	-0.144	-0.137	-0.141	-0.141
LNTIME×SP3	0.283	0.279	0.276	0.276
LNTIME2×SP1	-0.146**	-0.147**	-0.147**	-0.146**
LNTIME2×SP2	-0.00717	-0.00954	-0.0096	-0.00956
LNTIME2×SP3	-0.0867~	-0.0860~	-0.0854~	-0.0854~
MALE×SP1	2.110***	2.109***	2.109***	1.501***
MALE×SP2	0.405	0.0429		
MALE×SP3	0.665	0.452		
BLACK×SP1	3.080***	3.081***	3.080***	2.195***
BLACK×SP2	1.464*	0.468	0.600**	0.601**
BLACK×SP3	1.001	0.784	0.460~	0.461~
FRPL×SP1	2.877***	2.874***	2.877***	2.423***
FRPL×SP2	0.979~			
FRPL×SP3	0.485			
MALE×BLACK×SP1	-1.095~	-1.094~	-1.095~	
MALE×BLACK×SP2	-0.156	0.244		
MALE×BLACK×SP3	-0.406	-0.399		
MALE×FRPL×SP1	-1.453*	-1.451*	-1.452*	-0.949**
MALE×FRPL×SP2	0.0296			
MALE×FRPL×SP3	-0.231			
BLACK×FRPL×SP1	-1.738*	-1.736*	-1.738*	-0.929**
BLACK×FRPL×SP2	-1.247**			
BLACK×FRPL×SP3	-0.308			
MALE×BLACK×FRPL×SP1	0.975	0.973	0.974	
rho	0.025	0.024	0.024	0.024
N	596,789	596,789	596,789	596,789
rank	29	23	19	18
-2LL	8289.6	8297.4	8301.2	8304.2
LR Test (compared to Model 20)		7.835	11.641	14.654
df		6	10	12
p-value		0.25	0.31	0.261

*~ p<0.10, * p<0.05, ** p<0.01, *** p<0.001*

Table A5. Parameter estimates, approximate *p*-values, and goodness-of-fit statistics from five "final" fitted multiple-spell hazard models predicting the risk of out-of-school suspension by spell, time within spell, and critical covariates, specifying "robust" standard errors to account for correlations in risk for students attending the same district school.

	Model B-T1	Model D-T1	Model E-T1	Model F-T1	Model G-T1
SPELL1	-10.25***	-10.59***	-11.04***	-11.21***	-12.50***
SPELL2	-4.678***	-4.678***	-4.909***	-4.678***	-4.909***
SPELL3	-4.693***	-4.693***	-4.876***	-4.693***	-4.876***
LNTIME×SP1	1.601***	1.598***	1.315**	1.504***	1.260**
LNTIME×SP2	-0.0459	-0.0459	-0.127	-0.0459	-0.127
LNTIME×SP3	0.357	0.357	0.257	0.357	0.257
LNTIME2×SP1	-0.208***	-0.207***	-0.163***	-0.191***	-0.153***
LNTIME2×SP2	-0.0261	-0.0261	-0.0128	-0.0261	-0.0128
LNTIME2×SP3	-0.0985*	-0.0985*	-0.0824~	-0.0985*	-0.0824~
MALE×SP1		0.596***			1.495***
BLACK×SP1			2.358***		2.261***
BLACK×SP2			0.553**		0.553**
BLACK×SP3			0.428		0.428
FRPL×SP1				1.718***	2.434***
MALE×FRPL×SP1					-0.924**
BLACK×FRPL×SP1					-1.019**
N	798,960	798,960	596,789	798,960	596,789
rank	9	10	12	10	16
-2LL	10125.6	10096.2	8428.6	9908.0	8322.4

*~ p<0.10, * p<0.05, ** p<0.01, *** p<0.001*

Table A6. Parameter estimates, approximate *p*-values, and goodness-of-fit statistics from five "final" fitted multiple-spell hazard models predicting the risk of out-of-school suspension by spell, time within spell, and critical covariates, specifying additional parameter for each school (not shown) to account for correlations in risk for students attending the same district school.

	Model B-T2	Model D-T2	Model E-T2	Model F-T2	Model G-T2
SPELL1	-10.23***	-10.57***	-11.04***	-11.19***	-12.53***
SPELL2	-4.785***	-4.766***	-5.073***	-4.783***	-5.095***
SPELL3	-4.856***	-4.837***	-5.090***	-4.854***	-5.102***
LNTIME×SP1	1.588***	1.588***	1.282**	1.482***	1.214**
LNTIME×SP2	-0.0643	-0.0637	-0.148	-0.0588	-0.146
LNTIME×SP3	0.379	0.379	0.282	0.382	0.28
LNTIME2×SP1	-0.205***	-0.204***	-0.157***	-0.186***	-0.144**
LNTIME2×P2	-0.0214	-0.0215	-0.00794	-0.0226	-0.00834
LNTIME2×SP3	-0.101*	-0.101*	-0.0864~	-0.102*	-0.0861~
MALE×SP1		0.616***			1.507***
BLACK×SP1			2.321***		2.171***
BLACK×SP2			0.607**		0.610**
BLACK×SP3			0.474~		0.470~
FRPL×SP1				1.702***	2.412***
MALE×FRPL×SP1					-0.958**
BLACK×FRPL×SP1					-0.889**
N	798,960	798,960	533,329	798,960	533,329
-2LL	10041.4	10010.0	8382.2	9838.4	8273.8

*~ p<0.10, * p<0.05, ** p<0.01, *** p<0.001*

Table A7. Parameter estimates, approximate *p*-values, and goodness-of-fit statistics from five "final" fitted multiple-spell hazard models predicting the risk of out-of-school suspension by spell, time within spell, and critical covariates, and removing one school with the highest suspension rate from the analysis.

	Model B-T3	Model D-T3	Model E-T3	Model F-T3	Model G-T3
SPELL1	-9.573***	-9.906***	-10.28***	-11.27***	-11.77***
SPELL2	-5.138***	-5.144***	-5.153***	-4.844***	-5.163***
SPELL3	-5.686***	-5.691***	-5.681***	-4.906***	-5.683***
LNTIME×SP1	1.174**	1.175**	0.880*	1.485***	0.826*
LNTIME×SP2	0.0386	0.0391	-0.0233	-0.0559	-0.0215
LNTIME×SP3	0.742~	0.742~	0.622	0.378	0.62
LNTIME2×SP1	-0.156***	-0.156***	-0.112*	-0.187***	-0.101*
LNTIME2×P2	-0.0305	-0.0306	-0.0209	-0.0233	-0.0213
LNTIME2×SP3	-0.144**	-0.144**	-0.125*	-0.102*	-0.125*
MALE×SP1		0.580***			1.482***
BLACK×SP1			2.319***		2.196***
BLACK×SP2			0.334		0.336
BLACK×SP3			0.342		0.341
FRPL×SP1				1.710***	2.416***
MALE×FRPL×SP1					-0.962**
BLACK×FRPL×SP1					-0.907**
N	720,134	720,134	533,329	720,134	533,329
rho	0.041	0.043	0.018	0.033	0.02
rank	10	11	13	11	17
Log-likelihood	-4180.3	-4168.7	-3469.7	-4935.9	-3425

*~ p<0.10, * p<0.05, ** p<0.01, *** p<0.001*

Table A8. Parameter estimates, approximate *p*-values, and goodness-of-fit statistics for the five "final" fitted multiple-spell hazard models predicting the risk of out-of-school suspension by spell, time within spell, and free/reduced-price lunch status, and a proportional-odds specification of time, in school years, with Grade 6 as the reference category.

	Model B-T4	Model D-T4	Model E-T4	Model F-T4	Model G-T4
SPELL1	-7.315***	-7.652***	-8.480***	-8.334***	-9.955***
SPELL2	-4.969***	-4.977***	-5.150***	-4.980***	-5.188***
SPELL3	-4.403***	-4.412***	-4.907***	-4.428***	-4.944***
GRADE7	-0.620***	-0.614***	-0.479***	-0.578***	-0.442***
GRADE8	-1.009***	-1.001***	-0.749***	-0.949***	-0.702***
MALE×SP1		0.605***			1.463***
BLACK×SP1			2.292***		2.168***
BLACK×SP2			0.263		0.271
BLACK×SP3			0.515~		0.514~
FRPL×SP1				1.697***	2.398***
MALE×FRPL×SP1					-0.963**
BLACK×FRPL×SP1					-0.911**
N	798,960	798,960	533,329	798,960	533,329
rho	0.041	0.04	0.016	0.029	0.017
rank	6	7	9	7	13
Log-likelihood	-5027.5	-4161.8	-3462.4	-4925.7	-3419.3

*~ p<0.10, * p<0.05, ** p<0.01, *** p<0.001*

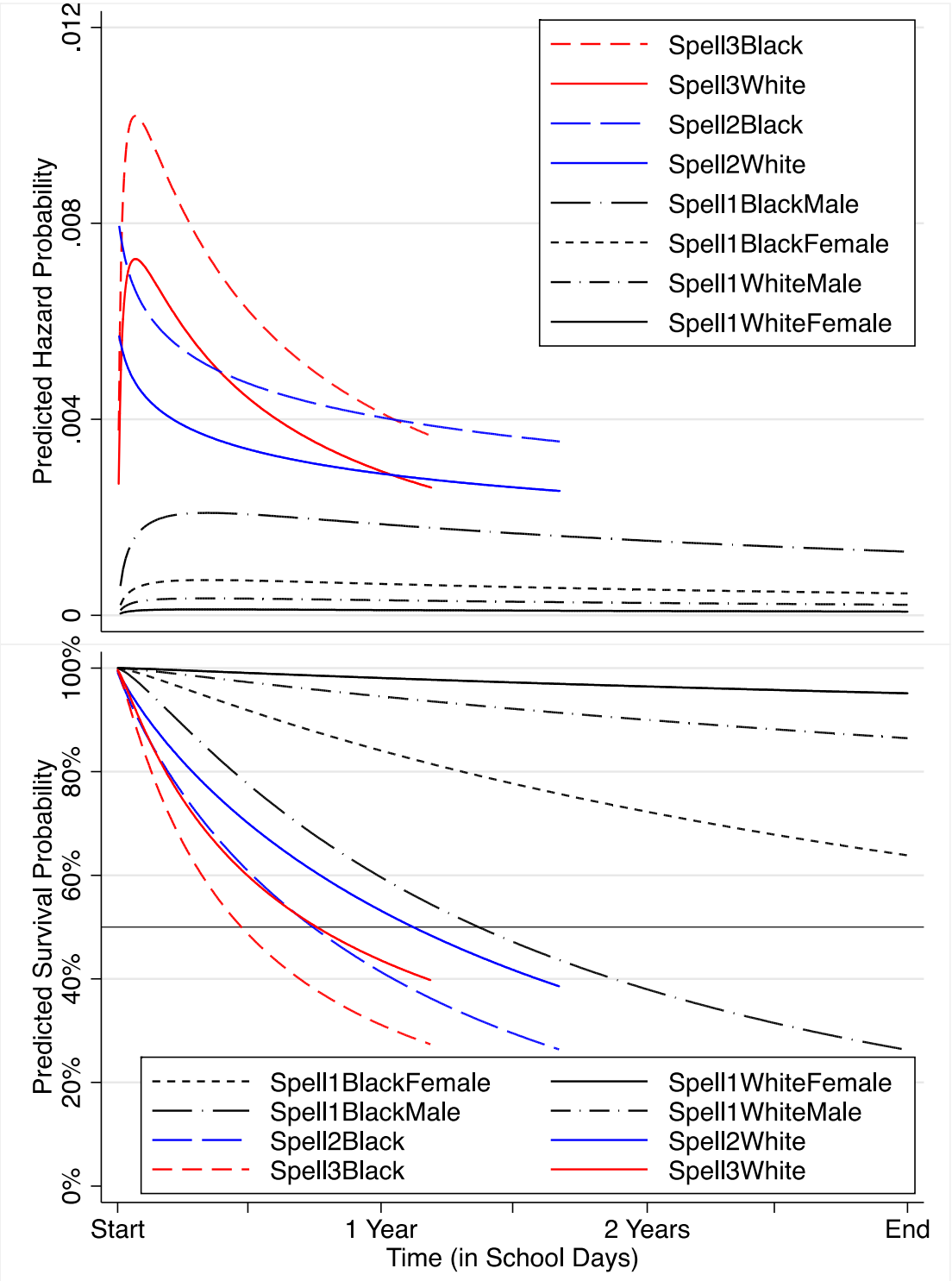

Figure A1. Predicted hazard probability (top panel) and survival probability (bottom panel) of a students' first, second and third suspension from middle school, Black students and White students, at average levels of free or reduced-price lunch status ($n_{students}$=1,274, $n_{suspensions}$=613). Predicted values obtained from fitting Model GT1 in Table A6.